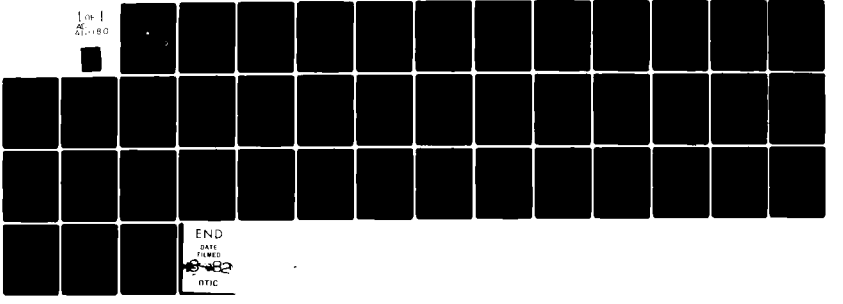
AD-A111 180 NAVAL BIODYNAMICS LAB NEW ORLEANS LA F/G 5/9
PERSPECTIVES IN PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL --ETC(U)
NOV 81 R S KENNEDY, A C BITTNER, M M HARBESON
UNCLASSIFIED NBDL-80R004 NL



NBDL - 80R004

PERSPECTIVES IN PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH (PETER): COLLECTED PAPERS

Robert S. Kennedy, Alvah C. Bittner, Jr., Marv M Harbeson
and Marshall B. Jones

November 1981

NAVAL BIODYNAMICS LABORATORY
New Orleans, Louisiana

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER NBDL-80R004	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Perspectives in Performance Evaluation Tests for Environemntal Research (PETER)		5. TYPE OF REPORT & PERIOD COVERED Research Report 6. PERFORMING ORG. REPORT NUMBER NBDL-80R004
7. AUTHOR(s) Robert S. Kennedy, Alvah C. Bittner, Jr., Mary M. Harbeson, and Marshall B. Jones		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Naval Biodynamics Laboratory Box 29407 New Orleans, LA 70189		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS MF58.524-002-5027
11. CONTROLLING OFFICE NAME AND ADDRESS Naval Medical Research and Development Command Bethesda, MD 20014		12. REPORT DATE November 1981 13. NUMBER OF PAGES 37
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release, distribution unlimited

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Human Performance Testing, Repeated Measurement, PETER

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The Performance Evaluation Tests for Environmental Research (PETER) program was begun at NBDL in 1977. This report includes four papers which were written between 1977 and 1980 describing progress and developments in this program. "An Engineering Approach to the Standardization of Performance Evaluation Tests for Environmental Research (PETER)" delineates the structure of the PETER paradigm, describes representative results and discusses implications of the results to previous and future research. "Assessing Productivity and Well-Being in Navy Workplaces" explains how Jones' rate-terminal performance and

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

BLOCK 20. ABSTRACT CONTINUED

theory of skill acquisition has been applied to the study of complex human performance and abilities. Examples from two tests administered under a fifteen day repeated measures paradigm are presented to illustrate the methodological approach employed in the PETER program. Application of these methods to selection and training research is suggested. "Progress in the Analysis of a Performance Evaluation Test for Environmental Research (PETER)" describes the preliminary results of ten tests which had been completed by October 1978. "The Development of a Navy Performance Evaluation Test for Environmental Research (PETER)" describes the earliest plan for developing PETER as it was formulated in 1977. It describes the philosophy and principles upon which the PETER program was based.

NBDL - 80R004

PERSPECTIVES IN PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH (PETER)

Robert S. Kennedy, Alvah C. Bittner, Jr., Mary M. Harbeson
and Marshall B. Jones

November 1981

Bureau of Medicine and Surgery
Work Unit No. MF58.524-002-5027

Approved by

Channing L. Ewing, M. D.
Chief Scientist

Released by

Captain J. E. Wenger MC USN
Commanding Officer

Naval Biodynamics Laboratory
Box 29407
New Orleans, LA 70189

SUMMARY PAGE

THE PROBLEM

Human performance testing in unusual environments such as ship motion and vibration almost always involves repeated testing of the same individuals. The purpose of the Performance Evaluation Tests for Environmental Research (PETER) program was to standardize a test battery for use in repeated measures experiments.

FINDINGS

The Performance Evaluation Tests for Environmental Research (PETER) program was begun at NBDL in 1977. This report includes four papers which were written between 1977 and 1980 describing progress and developments in this program. "An Engineering Approach to the Standardization of Performance Evaluation Tests for Environmental Research (PETER)" delineates the structure of the PETER paradigm, describes representative results and discusses implications of the results to previous and future research. "Assessing Productivity and Well-Being in Navy Workplaces" explains how Jones' rate-terminal theory of skill acquisition has been applied to the study of complex human performance and abilities. Examples from two tests administered under a fifteen day repeated measures paradigm are presented to illustrate the methodological approach employed in the PETER program. Application of these methods to selection and training research is suggested. "Progress in the Analysis of a Performance Evaluation Test for Environmental Research (PETER)" describes the preliminary results of ten tests which had been completed by October 1978. "The Development of a Navy Performance Evaluation Test for Environmental Research (PETER)" describes the earliest plan for developing PETER as it was formulated in 1977. It describes the philosophy and principles upon which the PETER program was based.

RECOMMENDATIONS

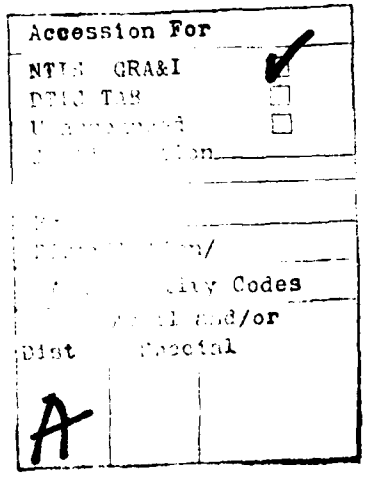
It is recommended that only stable and reliable tests be used in repeated measures experiments.

ACKNOWLEDGEMENTS

The authors are indebted to Dr. Channing L. Ewing, Ms. Michele Krause, Ms. Susan Jones, and Mr. Mike Shewmake for their contributions to the PETER program. Research on repeated measures testing is continuing at NBDL, but the name of the program has been changed to Performance Tests for Repeated Measures. Dr. Kennedy is now with the Canyon Research Group, 1040 Woodcock Road, Orlando, FL. 32803. Dr. Jones is with the Department of Behavioral Sciences, The Hershey Medical Center, Hershey, PA. 17033.

Trade names of materials or products of commercial or non-government organizations are cited where essential for precision in describing research procedures or evaluation of results. Their use does not constitute official endorsement or approval of the use of such commercial hardware or software.

TABLE OF CONTENTS



Each of these papers was presented at a professional meeting or symposium. Acknowledgement of previous publication appears at the beginning of each paper.

PROCEEDINGS OF THE 11TH ANNUAL CONFERENCE OF THE ENVIRONMENTAL
DESIGN AND RESEARCH ASSOCIATION (EDRA)
Charleston, South Carolina, 2-6 March 1980

AN ENGINEERING APPROACH TO THE STANDARDIZATION OF PERFORMANCE EVALUATION TESTS FOR ENVIRONMENTAL RESEARCH (PETER)

Robert S. Kennedy, Alvah C. Bittner, Jr., and Mary M. Harbeson
Naval Aerospace Medical Research Laboratory Detachment, New Orleans, LA

ABSTRACT

Many investigators have documented the problems of measuring performance in unusual environments. Reliable, valid, and standardized test batteries for repeated administrations have not been previously developed. This paper describes progress in developing such a battery: Performance Evaluation Tests for Environmental Research (PETER). In this program, the stability and sensitivity of performance tasks are studied over repeated sessions (15 days). The approach has been to test, at the same time of day, the same group of 20 healthy subjects in order to provide baselines and expected values. Thus far, 48 cognitive, perceptual and psychomotor tasks, mainly from the research literature, have been partially or completely evaluated. Subjecting these tasks to protracted practice reveals the following: (1) Most task performances do not asymptote, (2) most standard deviations are either homogeneous or they become regular, (3) and, more importantly, changes in reliabilities occur which cannot be anticipated from their means and standard deviations. The latter has not been commented upon before in this context. Based on these findings, it is believed that most previous environmental studies which employed a repeated measures paradigm should be seriously questioned or critically re-examined.

INTRODUCTION

An "engineering approach" to the development and standardization of the Performance Evaluation Tests for Environmental Research (PETER) battery has been previously proposed (Kennedy & Bittner, 1977). This engineering approach is directed at the test and evaluation (T&E) of performance tasks prior to their being employed for assessment of environmental effects. This T&E of performance tasks is similar to that which an engineer conducts to assess the stability of an instrument prior to its utilization. The goal of the PETER program is to study the possibly adverse effects of ship motion on performance. However, because PETER is being designed for repeated administrations, it will be directly applicable to studies in other environments and treatments (e.g., hyperbaric, thermal, drug).

TABLE 1

CATEGORIES IN THE STUDY OF ENVIRONMENTAL STRESS ON HUMANS

ADVERSE EFFECT CATEGORIES	DEFINITION
1. HEALTH AND SAFETY:	It exceeds medical limits adequate for safety and health.
2. COMFORT:	It is unpleasant, causes discomfort,
3. I/O QUALITY:	A physical aspect of the environment interacts to modify the input/output quality of stimulus or response.
4. CNS PROBLEMS:	It occasions major, identifiable changes in central nervous system (or "throughput") functioning.

Initially, PETER is being aimed at assessing central nervous system (CNS) functioning. CNS Problems as seen in Table 1 can be contrasted with other categories of "performance" decrements including: Health and Safety, Comfort, and Input/Output (I/O) Quality. Examples for each of these four categories appear in Table 2 for inertial environments and in Table 3 for hyperbaric. All of these categories are of concern to the individual who has the responsibilty for managing human effectiveness in a civilian or military setting, but each category implies a different type of performance degradation. The scientific and military literature rarely have distinguished between these categories. However, it is evident from inspection of Tables 1, 2, and 3 that specifying a category can imply the research strategy necessary for further study. Although present focus is on CNS Problems, future work will include the study of Comfort and I/O Quality Problems.

The purpose of this report is to delineate the structure of the PETER paradigm, describe representative results of the application, and discuss implications of the results to previous and future research.

TABLE 2

SHIP MOTION

ADVERSE EFFECT CATEGORIES	ILLUSTRATIVE PROBLEMS
1. HEALTH AND SAFETY:	Vomiting results in dehydration and accompanying problems.
2. COMFORT:	Nausea
3. I/O QUALITY:	
Input:	Movement of the platform may jiggle the image presented to the retina.
Output:	Body sway decreases limb steadiness.
4. CNS PROBLEMS:	
Idiopathic:	Soporific effects of motion
Nonidiopathic:	Estimates of the rate of passage of time have greater error during motion.

TABLE 3

HYPERBARIA

ADVERSE EFFECT CATEGORIES	ILLUSTRATIVE PROBLEMS
1. HEALTH AND SAFETY:	Aseptic necrosis
2. COMFORT:	Joint pain
3. I/O QUALITY:	
Input:	Chamber noise
Output:	Limb tremor
4. CNS PROBLEMS:	
Idiopathic :	High Pressure Nervous System Syndrome
Nonidiopathic:	Narcosis

THE PETER PARADIGM

Method

Task Selection. The strategy in PETER has been to consider tasks which purport to assess mental work. Initially, tasks which meet one or more of the following criteria are being selected for test and evaluation: (1) task performance has been reported to be disrupted in a thermal, or inertial or hyperbaric environment; (2) a concurrence in the scientific literature that some element of cognition, information processing, memory, etc., is being assessed by the task; or (3) the task distinguishes normal from brain damaged populations. This strategy is directed at obtaining a comprehensive selection of old and new tasks. In future studies, more real world oriented tasks will be examined along with these laboratory tasks.

Subjects. Twenty full time research subjects form the experimental population. These men are fit, average or above in intelligence, motivated to perform, and under constant military supervision and daily medical assessment (Thomas, Majewski, Ewing & Gilbert, 1977). All volunteer subjects were recruited and evaluated in accordance with procedures specified in Secretary of the Navy Instruction 3900.39 and Bureau of Medicine Instruction 3900.6. The instructions require voluntary informed consent and meet prevailing national and international guidelines.

Analysis. The test and evaluation plan is to obtain descriptive statistics for each test as it is performed for 15 workday mornings (8 - 10 AM). Analyses of means, standard deviations and correlations are used in the evaluation of tasks. Means over days and across subjects are analyzed to see whether they meet any of three criteria for mean stability: (1) plateau, or level across trials, (2) asymptotic, or approach to unchanging values after some point in training, or (3) slow, approximately linear increase, after some number of trials. Further, standard deviations across subjects are examined to see whether they are "stable" (i.e., constant) after some point in training. Lastly, cross trial reliabilities are studied to see whether they are "differentially stable", that is, have constant correlations with subsequent trials after some point in training (Jones, 1969, 1972). If criteria for the stability of the means, standard deviations and correlations are met, then a task can be recommended for tentative inclusion in the PETER battery. Ultimate inclusion in PETER will depend on factorial uniqueness and validity analyses which will be conducted in later stages of PETER development.

Rationale

The T&E approach described above was motivated by the pre, per, post (PPP), paradigm typically employed in environmental assessment research. The PPP paradigm assesses subjects for a number of trials: pre-exposure; during or per-exposure; and post-exposure. In these studies, small numbers of subjects, frequently less than six, are generally employed and simple repeated measures ANOVA are used to analyze the results. The PPP paradigm has many variants (e.g., addition of a nonexposure control group). However, whatever variant, the PPP paradigm has stringent requirements which must be met before results can be analyzed and interpreted meaningfully.

The criteria for mean, standard deviation, and correlation stability which are delineated above, must be met if the PPP paradigm is to be employed. In particular, changes in means over trials, other than slow, linear changes, can hide change due to an environment. For example, if means are changing over sessions when an environmental condition is encountered, it may not be determined whether it was overall level of performance which was disrupted or the learning. In addition, failure to meet either the standard deviation or reliability correlation requirements is equivalent to violating the compound symmetry assumptions of simple repeated measures ANOVA (Winer, 1972). Multivariate analysis methods might appear to offer an alternative to the simple ANOVA, however, these methods require substantially more subjects than trials (cf., Morrison, 1967). Additionally, the changing nature of what-is-being-measured, is signalled by differentially unstable reliability correlations (cf., Alvares & Hulin, 1972) which in turn makes attribution of effect difficult if not impossible (Bittner, 1979). Obviously, short of a major paradigm shift the stability criteria specified above must be met.

Differential stability of the reliability correlations is not the only feature to look at in evaluation of tasks for PETER. "Task definition", (Jones, 1979), the absolute magnitude of the reliability ($\bar{r}$) after stabilization is also considered. Unless task definition is substantial, sensitivity to differences between conditions may be poor. This may be seen on examination of Equation (1) which compares control and experimental condition means M_c and M_e, where the respective standard deviations are SD_c and SD_e, and where r_{ce} is the inter-trial correlation. With equal standard deviations, the standard error (1) may be seen to approach zero as the retest reliability approaches $r_{ce} = 1.00$. Conversely, the absence of reliability ($r_{ce} = 0$) implies that the size of the denominator (1) will be equivalent to the use of independent groups. Indeed, when the reliability is low, ($r < .40$) a few more subjects in each of two independent samples will result in more precision of the error term than is derived by repeated measures on the same subject. Caution should be employed in examinations of task definition and care should be taken to consider the time required to obtain a particular task datum. The Spearman-Brown adjustment (Allen & Yen, 1972, p.79) and similar approaches imply that increased data sampling will increase reliability hence, task reliability can be improved by increasing data collection time. Notwithstanding, task definition is employed in PETER but this must be tempered by consideration of the time required for taking measurements.

$$t = (M_c - M_e) / \sqrt{(SD_c^2 + SD_e^2 - 2r_{ce}SD_cSD_e)/N} \qquad (1)$$

RESULTS

Overview

Thus far, 48 tasks have been studied. Thirty of these have been completely analyzed and copies of the data can be obtained on request. The remainder are in various stages of completion with preprints available for 10 of them. The studied tasks tap functions from many areas of the human performance spectrum and have been drawn from a number of collections of tasks including: Rose (1974); Ekstrom, French, Harman & Derman (1976); Wechsler (1955) and others. It is suspected that as many as 200 total tests will eventually need to be evaluated in this way but a preliminary battery could be selected now on the basis of available findings. Results, reported below, will center around two tasks which are representative of those studied thusfar.

Representative Tasks

Air Combat Maneuvering. Figure 1 shows the means and standard deviations for the Air Combat Maneuvering (ACM) task (Jones, Kennedy & Bittner, in preparation) over 15 days. The means increase steadily through Day 14 with the increase being greatest during the first four days. Days 14 and 15 are the same. The standard deviations increase slightly through Day 5 and then remain constant through Day 15. Figure 2 is constructed from Table 4. Although correlations throughout the matrix are high ($r \rangle .70$) the earlier days (1, 2, & 4) are lower and more variable than later days. Base Day 6, 10 & 12 correlations are over .90 and remain constant with those following indicating differential stability.

Time Estimation. The means and standard deviations for the Time Estimation Test (McCauley, Kennedy, & Bittner, 1979) are shown in Figure 3. Both means and standard deviations appear approximately level throughout the experiment, with the standard deviation covarying with the small fluctuations of the mean. Figure 4, which was constructed from Table 5 shows the reliabilities of selected base days and those following for the Time Estimation Test. Although the reliabilities between adjacent days appear satisfactory, the reliabilities for Base Days 1 through 11 tend to decrease as a function of increasing days of separation, with correlations for the earlier days falling off more quickly and more dramatically. Correlations for Base Day 12 and those following were high ($\bar{r} = .85$) and a relatively shallow decrease in correlations with following days is seen.

TABLE 4

Air Combat Maneuvering Task (ATARI™ 1): Reliabilities Over 15 Days (N=13)

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.85	.77	.73	.88	.82	.79	.81	.77	.73	.72	.81	.76	.73	.77
2		.92	.87	.84	.83	.73	.82	.76	.70	.73	.77	.76	.74	.74
3			.90	.88	.84	.73	.80	.81	.70	.81	.73	.79	.74	.78
4				.88	.88	.84	.87	.86	.82	.91	.85	.89	.86	.86
5					.95	.91	.95	.94	.90	.93	.91	.93	.89	.92
6						.93	.97	.98	.92	.91	.94	.94	.94	.95
7							.97	.92	.93	.94	.96	.94	.93	.96
8								.95	.95	.93	.97	.94	.94	.96
9									.92	.94	.93	.94	.94	.94
10										.93	.98	.94	.93	.94
11											.93	.96	.94	.95
12												.95	.95	.96
13													.98	.98
14														.97

TABLE 5

Time Estimation: Constant Error (CE) Reliabilities Over 15 Days (n=19)

Days	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.80*	.40	-.14	.08	-.04	.16	.08	.03	-.12	-.19	-.05	-.21	-.26	-.24
2		.59	.22	.34	.28	.44	.40	.30	.14	.07	.16	-.05	-.02	-.07
3			.67	.73	.49	.54	.37	.20	.09	.12	.16	.12	.06	.03
4				.70	.69	.65	.53	.38	.28	.25	.27	.28	.19	.12
5					.80	.65	.62	.55	.38	.32	.42	.37	.36	.28
6						.83	.87	.82	.63	.57	.57	.52	.55	.37
7							.79	.70	.61	.53	.61	.53	.46	.39
8								.94	.80	.75	.72	.57	.66	.47
9									.84	.73	.72	.54	.62	.46
10										.76	.90	.82	.78	.78
11											.75	.61	.70	.54
12												.88	.84	.83
13													.89	.96
14														.90

FIGURES

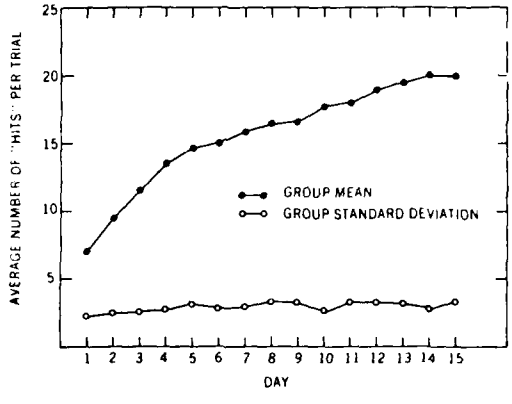


Figure 1. ACM (ATARI 1) means and standard deviations over 15 days (n=13).

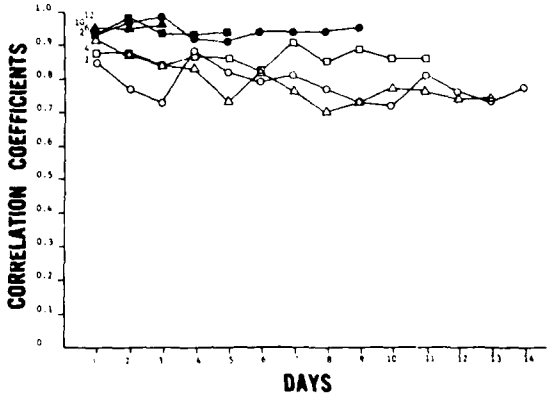


Figure 2. ACM (ATARI 1) reliabilities between selected base days (1, 2, 4, 6, 10, & 12 and those following over 15 days (n=13).

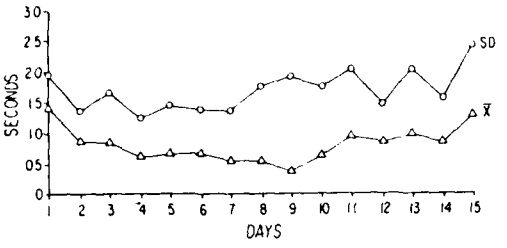


Figure 3. Time Estimation CE score means ($\bar{x}$) and standard deviations over 15 days (n=19).

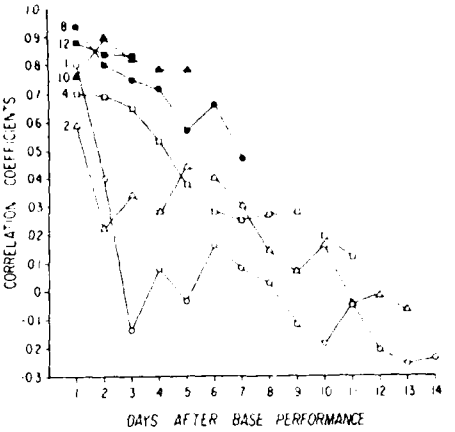


Figure 4. Time Estimation CE score reliabilities between selected base (SD) days (1, 2, 4, 8, 10 & 12) and those following over 15 days (n=18).

DISCUSSION

Changes in mean performance for the most part improve over the 15 days of an experiment for nearly all tasks studied by the PETER paradigm. Figure 1, ACM task from the Atari series of video games, is characteristic of what we find routinely, viz., a learning curve. Of note is that this test represents 30 minutes/day for three weeks, a lot of practice. Contrast this function with Time Estimation (Figure 3) where no learning curve is apparent, suggestive of a more desireable test from the standpoint of mean stability. In addition, a comparison of the standard deviations of both tasks show, if anything, greater stability for the Time Estimation task. However, comparison of Figures 2 and 4 which contain traces of correlation coefficients for these tasks tell a radically different story. Differential stability of the ACM task is obtained early and is of substantially greater magnitude than the marginally stable Time Estimation test. These two tests underscore the importance of the reliability - a neglected statistic in performance testing in adverse environments.

Not all tests behave similarly and, all combinations of mean and standard deviation changes can occur with or without stabilized correlations. In addition, results have shown that less than half of the tests which have been so studied (Jones, 1979) meet the criteria of stabilized reliability correlations. Of the ten tasks which have been reported, six tasks stabilize quickly and have acceptable task definition: Code Substitution (Wechsler, 1958), ACM from the Atari video game system (Jones, Kennedy & Bittner, in preparation), Grammatical Reasoning (Rose, 1974), Arithmetic (Seales, Kennedy & Bittner, 1979), Stroop Color-Words (Harbeson, Kennedy & Bittner, 1979), and Two-Dimensional Tracking (Damos, 1979). Critical Tracking (Damos, Kennedy & Bittner, 1979) also stabilizes with acceptable task definition but findings are less clear cut. Arithmetic is best in magnitude and quickness of correlational stability and ACM is next best. Four tasks: Complex Counting, (Kennedy & Bittner, 1979), Time Estimation, Letter Search, and the Spoke Trail-Making Test (Kennedy & Bittner, 1978) either do not stabilize or, if they do, have unacceptably low task definition.

In conclusion, half of the tests we have studied lack differential stabilization as revealed by examining the correlations. Given that this result occurs in tasks which may have stable means and standard deviations and were largely drawn from established batteries, it might be conjectured that of all previous investigations in adverse environments, many may have been conducted employing unstable tasks. Differential stability, as discussed earlier, is required for valid and meaningful analysis. It is believed that when the results of environmental studies have been based on tasks not shown as differentially stable or employing independent groups designs, these studies should be seriously questioned or critically re-examined.

REFERENCES

Alvares, K. M. & Hulin, C. L. Two explanations of temporal changes in ability-skill relationships: A literature review and theoretical analysis. Human Factors, 1972, 14, 295-308.

Allen, M. J. & Yen, W. M. Introduction to Measurement Theory, Monterey, CA: Brooks Cole Publishing Company, 1979.

Bittner, Jr., A. C. Statistical tests for differential stability. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston, October, 1979.

Damos, D. L. The stability of tracking tasks performed singly and in dual modes. Paper presented at the ONR Contractors' Meeting on Information Processing Abilities, New Orleans, LA, 20-23 February 1979.

Damos, D. L., Kennedy, R. S. & Bittner, Jr., A. C. Development of Performance Evaluation Tests for Environmental Research (PETER): Critical tracking test. Proceedings of the 50th Annual Meeting of the Aerospace Medical Association, Washington, D.C., May, 1979. (AD A066719)

Ekstrom, R. B., French, J. W., Harman, H. H. & Derman, D. Manual for kit of factor-referenced cognitive tests. Princeton, N.J.: Educational Testing Service, 1976.

Harbeson, M. M., Kennedy, R. S., & Bittner, Jr., A. C. A comparison of the Stroop Test to other tasks for studies of environmental stress. Proceedings of the 12th Annual Meeting of the Human Factors Association of Canada, Bracebridge, Ontario, Canada, 6-8 September, 1979.

Jones, M. B. *Differential processes in acquisition.* In E. A. Bilodeau and I. McD. Bilodeau (Eds.), Principles of skill acquisition. New York: Academic Press, 1969.

Jones, M. B. Stabilization and Task Definition in a Performance Test Battery. Final report on Contract No. N0023-79-M-5089 with the U. S. Naval Aerospace Medical Research Laboratory Detachment, New Orleans, Louisiana, 12 May 1979.

Jones, M. B. Individual differences. In R. N. Singer (Ed.). The psychomotor domain. Philadelphia: Lea and Febiger, 1972.

Jones, M. B., Kennedy, R. S., & Bittner, Jr., A. C. A video game for performance testing. (in preparation)

Kennedy, R. S. & Bittner, Jr., A. C. Progress in the analysis of Performance Evaluation Tests for Environmental Research (PETER). Proceedings of the 22nd Annual Meeting of the Human Factors Society, Detroit, October, 1978. (AD A060676)

Kennedy, R. S. & Bittner, Jr., A. C. The development of a Performance Evaluation Test for Environmental Research (PETER). In, Productivity enhancement in Navy systems. San Diego, California: Naval Personnel and Development Center, October, 1977.

Kennedy, R. S. & Bittner, Jr., A. C. Development of Performance Evaluation Tests for Environmental Research (PETER): Complex counting test. Aviation, Space and Environmental Medicine. (in press)

McCauley, M. E., Kennedy, R. S. & Bittner, Jr., A. C. Development of Performance Evaluation Tests for Environmental Research (PETER): Time estimation test. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston, October, 1979.

Morrison, D. F. Multivariate statistical methods. New York: McGraw Hill, 1967.

Rose, A. M. Human Information Processing: An Assessment and Research Battery. Doctoral Dissertation, Ann Arbor, MI: University of Michigan, 1974, (also published as AFOSR-PR-74-1372). AD-785-411.

Seales, D. M., Kennedy, R. S. & Bittner, Jr., A. C. Development of Performance Evaluation Tests for Environmental Research (PETER): Arithmetic computation. Proceedings of the 23rd Annual Meeting of the Human Factors Society, Boston, October, 1979.

Thomas, D. J., Majewski, P. L., Ewing, C. L. & Gilbert, N. S. Medical Qualification Procedures for Hazardous-duty Aeromedical Research. (Conference Proceedings No. 231, A3, pp. 1-13, 1978) London: AGARD, 1977.

Wechsler, D. The Measurement and Appraisal of Adult Intelligence. Baltimore: The Williams and Wilkins Co., 1958.

Winer, B. J. Statistical principles in experimental design (Second Edition). New York: McGraw Hill, 1971.

PROCEEDINGS OF THE 13TH ANNUAL MEETING OF THE HUMAN FACTORS ASSOCIATION OF CANADA, POINT IDEAL, ONTARIO, CANADA
SEPTEMBER 1980

ASSESSING PRODUCTIVITY AND WELL-BEING IN NAVY WORKPLACES

Robert S. Kennedy[1], Marshall B. Jones[2], and Mary M. Harbeson[3]
Naval Biodynamics Laboratory, New Orleans, LA[1,3]
Pennsylvania State University, Hershey, PA[2]

ABSTRACT

When individuals are required to work in arduous environments, such as may be encountered aboard ship, productivity and well-being can be reduced. The Performance Evaluation Tests for Environmental Research (PETER) battery is being designed to monitor the effects of such unusual environments. In the PETER program, Jones' rate-terminal theory of skill acquisition is being applied to the study of complex human performance and abilities. This model was originally derived from studies of motor skill acquisition and permits isolation of performance into two elements, one relating to the acquisition stage of training and the other to the capacity of the individual. The reliability of most test batteries has been determined over only two or three administrations, which assumes that stable (unchanging) abilities are being measured. Task performance, however, generally changes with practice. Unless, therefore, a task has been practiced until between-subject differences cease to change, it cannot be used reliably to measure environmental (or any other) effects. During the early trials on a test, subjects improve at different rates and, after extended practice, arrive at different terminal levels of skill. If subjects are tested for environmental effects during the acquisition phase, it is not possible to tell whether differences in performance are due to individual differences or to differences in exposure and transfer. It is only when a test is stable, that is, when mean performance levels off and the rank order of subjects ceases to change, that a test can measure environmental effects. Findings from the sixty tests which have been administered in a fifteen day repeated-measures paradigm support the rate-terminal theory of skill acquisition. Examples from two of these tests are presented to illustrate the methodological approach we employ for the study of complex mental functions. The application of these methods to selection and training research is suggested, and the critical re-examination or reinterpretation of human performance studies which have not taken repeated measures problems into consideration is recommended.

INTRODUCTION

Environmental stressors which are experienced in Navy workplaces, such as aboard ship, may reduce well-being and productivity. The gross effects of such arduous environments are readily observable, but in order to detect subtle effects a sensitive measuring instrument is necessary. Such a testing device could be used to predict the onset of decrements in performance, to select resistant personnel or to explore the possibility of training people to become more resistant. The Performance Evaluation Tests for Environmental Research (PETER) battery, which is being developed primarily to study ship motion, is being designed to be sensitive to subtle changes in performance. It is our opinion that this type of sensitivity has not been achieved in past human performance studies because adequate attention has not been given to the effects of practice.

Several years ago, Jones (1970a, 1970b) proposed a two process theory to describe the acquisition of motor skills. The theory posited an acquisition phase, in which persons improve at different rates and a terminal phase in which persons reach or approximate their individual limits. The theory therefore specifies (and experimental data support) that different persons begin at different points initially and arrive at different final values via different pathways. The theory further implies that, to the extent that the terminal process is reached, persons will cease to change positions relative to each other despite additional practice. In other words, several individuals may approach a task with differing experience levels and capacities, both of which influence their initial scores*. As practice continues, previous experience will begin to contribute proportionately less to a person's score, and individual differences in learning, or the readiness with which a person acquires his best performance, begins to influence his test score more. As the amount of experimental time increases proportional to previous practice, and as learning progresses, differences between subjects will become more attributable to actual differences in underlying ability, or capacity until finally , the amount of ability is largely what governs performance scores. Thus, an inter-session correlation matrix would present a distinctively different appearance if performance early versus late in practice was examined. Early in practice one would ordinarily observe the superdiagonal form (Jones, 1969) in which correlations between adjacent trials would be higher than comparisons which are more remote. Secondly, correlations of immediately adjacent trials (e.g., 1,2; 2,3;...) would be higher later (e.g., trials 10,11) rather than earlier (e.g., trials 2,3) in practice. Late in practice, if the theory holds, the correlation coefficients would become constant if the terminal process is reached so that no systematic differences would be present in the matrix as a function of temporal separation. If the terminal process is not reached, then the matrix will continue to show superdiagonal form (Jones, 1969). This concept is important for statistical as well as theoretical reasons. Repeated measures analysis of variance requires symmetry of the variance-covariance matrix and if learning is not accomplished during pretesting then systematic changes as described above can make interpretation of data using an ANOVA model (Winer, 1971; Morrison, 1967) difficult or impossible. Therefore the rate-terminal process theory provides theoretical underpinning for a statistical requirement. Moreover, it provides a way of looking at the results in order to determine whether stability of performance is attained.

The PETER program was begun to standardize a performance test battery in order to study the effects of adverse environments on humans (Kennedy & Bittner, 1977). It is desireable that the tests in the battery assess complex mental abilities which could be related as elements of Navy jobs. A natural consequence of research in this area of environmental stress is that

This research was performed under Navy Contract No. ME58,524,002-5027. The opinions are those of the authors and do not necessarily reflect those of the Department of the Navy.

*It is recognized that individual differences in motivation can add consequences but are ignored for this discussion.

generally each subject serves as his own control over many sessions. In other words, repeated measures analysis of variance is required. Moreover, within the context of the Jones' theory, performance on all tasks within the battery should be at terminal levels before an experimental treatment is introduced, in order that the changes which occur may be correctly and differentially attributed to the faculty or ability being tested. To our knowledge, no battery of performance tasks exists which would permit this inference to be made. Many batteries of primary mental abilities have been developed and most have been factor analyzed (cf. Carter, Kennedy, & Bittner, 1980a, for a review). None of these has been examined in terms of stability of subtests over sessions*, and generally the factor analyses which were performed were conducted on at most two replications. Recently, reviews of mean performance changes on WAIS (Thompson, 1975) and SAT (Nader Releases ETS Report, 1980) repeated testings have suggested that these tasks also may be less stable than previously considered. In WAIS, SAT, and factor analyzed batteries' reports, cross-session correlations of subtests are ordinarily not reported for more than 2 or 3 sessions. These issues bear directly on the standardization of a performance test battery for studying environmental stress; because stable mental abilities as well as stable performance skills will need to be measured in such a battery. Stability can only be determined empirically by testing over sessions. Thus, the question arises as to whether the rate-terminal process theory would provide a useful framework in which to evaluate the suitability of tests of simple and complex mental work. Specifically, do people exhibit differential rate processes when faculties such as short term memory (Sternberg, 1966), grammatical reasoning (Baddeley, 1968), or visualization (Ekstrom, French, Harman, & Derman, 1976) are tested, in the same way that they acquire the skill of turning a crank or pushing a lever (Jones, 1969).

METHOD

The PETER paradigm which has been described in detail elsewhere (Harbeson, Kennedy, & Bittner, 1979; Kennedy, Bittner, & Harbeson, 1980; Kennedy, Carter, & Bittner, 1980) entails testing approximately 20 persons, usually 15 minutes a day each, for 15 days on a series of tests of skills and abilities. Group means and standard deviations between subjects, and cross-session correlations are examined to determine whether they meet set criteria. The tests under study for potential inclusion in PETER are selected on the basis of meeting one or more of the following criteria: (a) the test appears in a factor analyzed battery, (b) the test measures an information processing construct supported by a body of research, (c) performance on the test has been experimentally disrupted in an adverse environmental condition of interest to the Navy (viz., motion, thermal, pressure), (d) the task taps a factor related to Navy jobs, or (e) the test is intrinsically motivating (cf. Carter et al. 1980a for additional information).

RESULTS

Thusfar sixty tests have been examined for stability. A preliminary report covering fifteen has been presented elsewhere (Kennedy, Carter, & Bittner, 1980). What follows are examples of two tasks which make qualitatively different demands of subjects. One, Grammatical Reasoning (Carter, Kennedy, & Bittner, 1980b) is a cognitive test, and the other a video game, Air Combat Maneuvering (Jones, Kennedy, & Bittner, 1980, in press) is largely a psychomotor task. Figure 1 shows mean and standard deviation performances for the Air Combat Maneuvering task. It may be seen that, typical of learning curves, the means increase dramatically over the first few (five) sessions, and that the rate of improvement becomes constant thereafter. Table 1 contains the cross-session correlations for this test. It is considered representative of the motor skill tasks examined thusfar and follows the generic descriptions of Jones (1980). Early in practice, the correlations degrade along each row, but later in practice (viz., in this case, after Day 6) the correlations appear symmetrical. That is, comparisons 6 days apart, (i.e., between Days 14 and 8) are the same as those close together (viz., between Days 13 and 14). Note also that the superdiagonal form (Jones, 1969) is absent after Day 6. Figure 2 shows data we consider representative of the cognitive tests that we have studied. The group means and between-subject standard deviations for Grammatical Reasoning also show a learning curve, and Table 2 shows similar form but lower correlations (e.g., task definitions) than Table 1**. Day 15, the last day, contains anomalous results, a common finding in our 15 day paradigm. Discounting Day 15, symmetrical correlations appear by Day 6 in Table 2 and are comparable to those shown in Table 1. The reasons for these systematic changes in correlation matrices are now described. Figure 3 shows a scatter plot of individual scores for the 23 subjects tested over the 15 days on Grammatical Reasoning. The overall impression is of a learning curve. Four different time-course performances were exhibited by these subjects, and they are separated into classes in Figures 4-7. Figure 4 illustrates subjects whose scores over the 15 sessions were essentially constant. Figure 5 shows subjects who improve with practice but all at the same rate. Figure 6 demonstrates subjects whose terminal level is correlated with initial level but the individuals appear to improve at different rates. Figure 7 reflects the full complexity of the two process theory whereby individual differences exist for initial and terminal levels as well as for the rates of learning.

DISCUSSION

Figure 7 is typical of the general findings of many of our experiments. That these outcomes would emerge for motor skill acquisition tasks was not surprising. However, that tests of information processing and tests of cognitive abilities would follow superdiagonal form has not been commented upon previously to our knowledge. These findings have profound implications not only for experiments into adverse environments but also for all other studies which follow a repeated measures design and where systematic change in cross-session correlations may occur. While time course changes similar to those of Figures 4, 5, and 6 are available in our work, they are the exception. On the other hand, the following exhibit data like Figure 7: Digit Span (McCafferty, Bittner, & Carter, 1980), Code Substitution (Pepper, Kennedy, Bittner, Wiker, 1980), Copying (Moran, Kimble & Mefferd, 1964) Letter Rotation and other mental ability tests; Letter Search

*An exception may be the Alluisi and Chiles (1967) battery which may have been subjected to a stability analysis during its early development. However, the inter-session and intertask reliabilities have not been reported to our knowledge.

**When these two tests are normalized (using a 5 minute base) for their disparite test lengths, ACM correlations are slightly poorer than Grammatical Reasoning.

(Rose, 1974), Item Recognition (Carter, Kennedy, Bittner, & Krause, 1980), and other information processing tasks; Free Recall, Running Recognition, and other memory tests (Harbeson, Krause, & Kennedy, 1980). Not surprisingly, the following psychomotor tests also show superdiagonal form over various periods of a 15 day experimental paradigm: Critical Tracking (Damos, Kennedy, & Bittner, 1979) Trail Making (Kennedy, Bittner, & Einbender, 1980) as well as several in a family of video games (Jones, et al. 1980a, in press). Arithmetic (Seales, Kennedy, & Bittner, 1980) shows data like Figure 5 and is an example of a test which stabilizes early. Only two studies from our data provide examples which resemble Figure 4 and these, Time Estimation (McCauley, Kennedy, & Bittner, in press) and Complex Counting (Kennedy & Bittner, 1980) show late if any stabilization of the correlations, possibly because knowledge of results is not provided in those tests. Compensatory tracking (Damos, Kennedy, & Bittner, 1980, in press) exhibits high correlations between initial and terminal ($r => .80$) performance (cf. Figure 6) but as many persons reach their best performance early (< 30 trials) as late (> 70 trials) in practice.

We feel that the implications of our findings are best viewed by references to the illustrations shown in Figures 4-7. For example, when Factor Analytic Studies of primary mental abilities are conducted by others using large samples on several paper and pencil tests but over only 1 and 2 administrations, it is implicit that time course changes in individual performance follow either Figure 4 or at least Figure 5. Yet, our data strongly suggest that tests which now appear in factor analyzed batteries often do not stabilize until after several administrations. This means that in previous factor analyses, the "primary mental ability" which emerged may have been complicated by individual differences in learning the "primary mental ability" test.

Implications similarly exist for Selection and Training Research where scores are used to predict subsequent performance. In these cases, it is essential that the initial scores be stable attributes of an individual because the test-retest reliability of a selection test score (e.g. spatial apperception) is the expected upper limit of the correlation of that score with an external criterion (e.g. success as an aviator). Thus, Selection and Training Research hopes for outcomes like Figure 4 but admits of outcomes like Figure 5 or 6. However, to the extent that Figure 7 can be expected to occur on selection tests and during training regima, inefficiency in prediction will result if predictor scores are unstable when related to a criterion. Obtaining stable test scores will assuredly improve predictive validity. We feel that implications also exist for Experimental Psychology, particularly for information processing and perception studies. When that discipline employs repeated measures designs, it often attempts to control the exposure history of subjects (or counterbalance) to account for sequence effects. Our data show that far more practice than is usually provided is necessary for stability. When inferences about particular hypothetical constructs (e.g. amphetamine modifies short term memory) are to be made, it is necessary to have the subjects' stable performances of short term memory (i.e. terminal process) separated from their acquisition (i.e. rate process) prior to the application of the experimental treatment. Only in this way may the experimental outcome be properly referred to the effect of the experimental treatment on the hypothetical construct. Otherwise, individual differences in acquisition are contaminated with the individual differences in the ability. It is our opinion that since Figure 7 appears to better reflect reality, research in the aforementioned fields of psychology should reexamine findings with this in mind. Moreover, it is our view that studies which report mean differences for asymptotic performances between ages, sexes and races, should also determine whether individual learning curves are similarly shaped or consider the possibility that subjects are merely following different paths to stability. It is possible that practice (previous experience) would account for proportionately more differences in performance than the different basic abilities of the groups. If so, there could be dramatic practical advantages. For example, training persons with poorer ability may result in greater increases in performance at less cost than selecting persons with high ability initially.

A test battery such as PETER could serve as a useful tool in assessing the effects of the work environment. Each individual would practice to asymptote on tests of various skills and abilities, and subsequently be tested in the work environment. Thus, it would be possible to determine subtle changes in performance for a particular individual, or for a particular function of that individual. Such a test battery could be used to monitor the daily effects of a hazardous environment, in which individuals were working, or for research on the environment. The results of such testing could be used as a warning to remove workers from dangerous conditions, or to select resistant workers, or to redesign the workplace.

CONCLUSION

In conclusion, in research on human performance, it is important to consider practice effects. If subjects are tested during the acquisition phase of training, it is not possible to tell whether differences in performance are due to individual differences, or are caused by the variable being studied. It is only when a test is stable that is, when mean performance levels off and the rank order of subjects ceases to change, that a test can be used as an accurate measuring device.

REFERENCES

Alluisi, E. A., & Chiles, W. D. Sustained performance, work-rest scheduling, and diurnal rhythms in man. Acta Psychologica, 1967, 27, 436-442.

Baddeley, A. D. A 3 minute reasoning test based on grammatical transformation. Psychonomic Science, 1968, 10, 341-342.

Carter, R. C., Kennedy, R. S., & Bittner, Jr., A. C. Selection of Performance Evaluation Tests for Environmental Research. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, 1980. (a)

Carter, R. C., Kennedy, R. S., & Bittner, Jr., A. C. Grammatical reasoning: A stable performance yardstick. Manuscript submitted for publication, 1980(b).

Carter, R. C., Kennedy, R. S., Bittner, Jr., A. C., & Krause, M. Item recognition as a performance evaluation test for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, 1980.

Damos, D. L., Kennedy, R. S., & Bittner, Jr., A. C. Development of a performance evaluation test for environmental research (PETER): Critical tracking test. Proceedings of the 50th Annual Scientific Meeting of the Aerospace Medical Association, 1979, 21.1-21.9.

Damos, D. L., Kennedy, R. S., & Bittner, Jr., A. C. The stability of tracking task performed singly and in dual modes. In preparation.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. Manual for kit of factor-referenced cognitive tests. Princeton, New Jersey: Educational Testing Service, 1976.

Harbeson, M. M., Kennedy, R. S., & Bittner, Jr., A. C. A comparison of the Stroop test to other tasks for studies of environmental stress. Proceedings of the 12th Annual Meeting of the Human Factors Association of Canada. Bracebridge, Ontario, 1979, 21.1-21.9.

Harbeson, M. M., Krause, M., & Kennedy, R. S. Comparison of memory tests for environmental research. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, 1980.

Harbeson, M. M., Krause, M., Kennedy, R. S., & Bittner, Jr., A. C. The Stroop as a Performance Evaluation Test for Environmental Research. Manuscript submitted for publication, 1980.

Jones, M. B. Differential processes in acquisition. In E. A. Bilodeau & I. McD. Bilodeau (Eds.), Principles of skill acquisition. New York: Academic Press, 1969, 141-170.

Jones, M. B. Rate and terminal processes in skill acquisition. American Journal of Psychology, 1970, 83, 222-236. (a)

Jones, M. B. A two-process theory of individual differences in motor learning. Psychological Review, 1970, 77, 353-360. (b)

Jones, M. B. Stabilization and task definition in a performance test battery. (NBDL Monograph No. M-0001) New Orleans, LA: Naval Biodynamics Laboratory, 1980.

Jones, M. B., Kennedy, R. S., & Bittner, Jr., A. C. Video games and convergence or divergence with practice. Proceedings of the Seventh Psychology in the DOD Symposium, USAF Academy, Colorado Springs, CO, 16-18 April 1980.

Jones, M. B., Kennedy, R. S., & Bittner, Jr., A. C. A video game for performance testing. American Journal of Psychology, in press.

Kennedy, R. S., & Bittner, Jr., A. C. The development of a Navy Performance Evaluation Test for Environmental Research (PETER). In L. T. Pope & D. Meister (Eds.), Productivity Enhancement: Personnel Performance Assessment in Navy Systems. Symposium presented at the Naval Personnel R & D Center, San Diego, October, 1977, 393-408. (NTIS No. AD A045047).

Kennedy, R. S., & Bittner, Jr., A. C. Development of performance evaluation tests for environmental research (PETER): complex counting. Aviation, Space, and Environmental Medicine, 1980, 51, 142-144.

Kennedy, R. S., Bittner, Jr., A. C., & Einbender, S. W. Development of performance evaluation tests for environmental research (PETER): trail making test. Unpublished manuscript, 1980.

Kennedy, R. S., Bittner, Jr., A. C., & Harbeson, M. M. An engineering approach to the standardization of performance evaluation tests for environmental research (PETER). Proceedings of the 11th Annual Conference of the Environmental Design Research Association (EDRA), Charleston, SC, 2-6 March, 1980.

Kennedy, R. S., Bittner, Jr., A. C., & Jones, M. B. The utility of commercially available television-computer games for assessing performance and other applications. Proceedings of the 51st Annual Scientific Meeting of the Aerospace Medical Association, 1980.

Kennedy, R. S., Carter, R. C., & Bittner, Jr., A. C. A catalogue of Performance Evaluation Tests for Environmental Research. Proceedings of the 24th Annual Meeting of the Human Factors Society, Los Angeles, 1980.

McCafferty, D. B., Bittner, Jr., A. C., & Carter, R. C. Performance evaluation tests for environmental research (PETER): Auditory digit span. Proceedings of the 24th Annual Meeting of the Human Factors Society, 1980.

McCauley, M. E., Kennedy, R. S., & Bittner, Jr., A. C. Development of performance evaluation tests for environmental research (PETER): time estimation test. Perceptual and Motor Skills, in press.

Moran, L. J., Kimble, J. P., & Mefferd, R. B. Repetitive psychometric measures: Equating alternate forms. Psychological Reports, 1964, 14, 335-338.

Morrison, D. F. Multivariate statistical methods. New York: McGraw-Hill, 1967.

Nader releases ETS report, hits tests as poor predictors of performance. APA Monitor, February, 1980, pp 1; 7.

Pepper, R. L., Kennedy, R. S. Bittner, Jr., A. C., & Wiker, S. F. Performance evaluation tests for environmental research (PETER): Code substitution test. Proceedings of the 7th Psychology in the DOD Symposium, USAF Academy, Colorado Springs, 1980.

Rose, A. M. Human information processing: An assessment and research battery. Human Performance Center, Technical Report No. 46, Ann Arbor: University of Michigan, 1974.

Seales, D. M., Kennedy, R. S., & Bittner, Jr., A. C. Development of performance evaluation tests for environmental research (PETER): arithmetic computation. Perceptual and Motor Skills, in press.

Sternberg, S. High speed scanning in human memory. Science, 1966, 153, 652-654.

Thompson, C. The effects of practice on intelligence tests. Unpublished manuscript, Wake Forest University, 1975.

Underwood, B. J., Boruch, R. F., & Malmi, R. A. The composition of episodic memory. (ONR Contract No. N00014-76-C-0270) (NTIS No. AD A040696).

Winer, B. J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.

TABLES

Table 1

Cross-session Correlations for Air Combat Maneuvering Test Over 15 Days (n = 22)

Day	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.78	.68	.59	.73	.72	.64	.70	.62	.61	.59	.64	.56	.65	.69
2	-	.87	.79	.81	.77	.68	.74	.70	.69	.67	.71	.69	.69	.72
3		-	.86	.86	.84	.72	.76	.81	.72	.81	.74	.79	.75	.78
4			-	.87	.86	.83	.84	.94	.82	.85	.86	.89	.86	.84
5				-	.93	.90	.91	.88	.92	.89	.89	.90	.91	.91
6					-	.91	.94	.93	.92	.91	.92	.90	.94	.93
7						-	.97	.90	.93	.93	.95	.93	.92	.95
8							-	.90	.93	.90	.94	.90	.93	.94
9								-	.89	.91	.92	.92	.89	.93
10									-	.91	.96	.92	.93	.93
11										-	.92	.93	.91	.93
12											-	.95	.91	.94
13												-	.93	.96
14													-	.96

Table 2

Cross-session Correlations
Grammatical Reasoning Test Over 15 Days (n = 23)

Day	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	.54	.70	.61	.56	.68	.60	.51	.65	.48	.60	.50	.54	.60	.31
2	-	.78	.67	.63	.66	.56	.64	.63	.43	.52	.57	.46	.70	.32
3		-	.88	.74	.88	.65	.78	.79	.67	.63	.73	.64	.76	.37
4			-	.78	.86	.89	.77	.76	.65	.68	.66	.58	.77	.43
5				-	.86	.82	.82	.86	.76	.82	.86	.83	.91	.68
6					-	.83	.85	.83	.79	.76	.76	.80	.86	.61
7						-	.82	.79	.83	.77	.77	.85	.90	.67
8							-	.86	.77	.77	.84	.80	.89	.68
9								-	.77	.88	.87	.83	.87	.69
10									-	.64	.80	.95	.79	.77
11										-	.82	.83	.84	.65
12											-	.89	.86	.67
13												-	.85	.73
14													-	.74

FIGURES

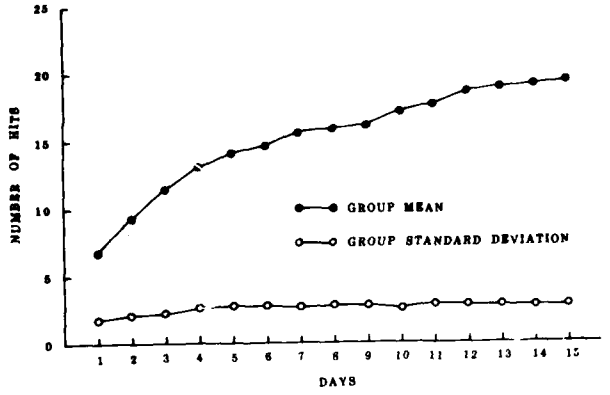


Figure 1. Means and standard deviations for Air Combat Maneuvering Test over 15 days (N=22).

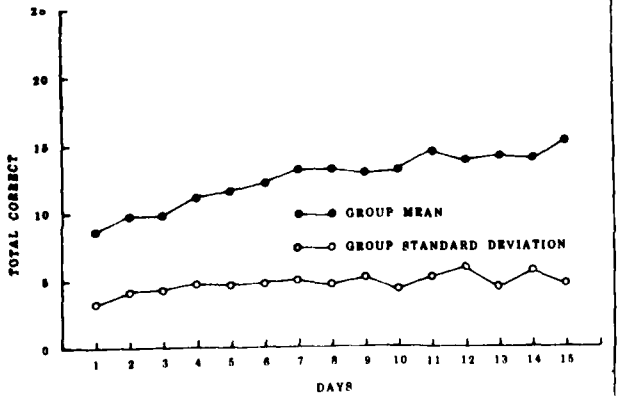


Figure 2. Means and standard deviations for Grammatical Reasoning Test over 15 days (N=23).

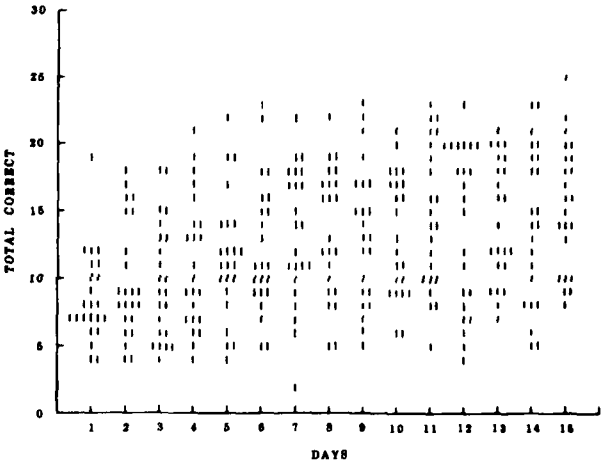


Figure 3. Scatter plot of total correct scores for 23 subjects over 15 days on the Grammatical Reasoning Test.

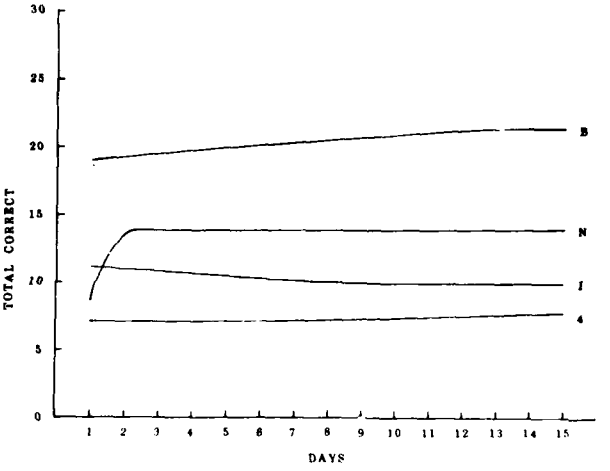


Figure 4. Individual learning curves of 4 subjects (4, B, I, N) whose scores appear constant over 15 days on the Grammatical Reasoning Test.

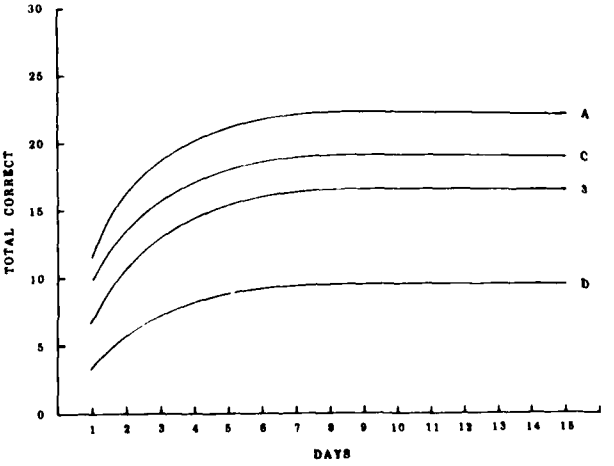


Figure 5. Individual learning curves of 4 subjects (3, A, C, D) who improve with practice at the same rate over 15 days on the Grammatical Reasoning Test.

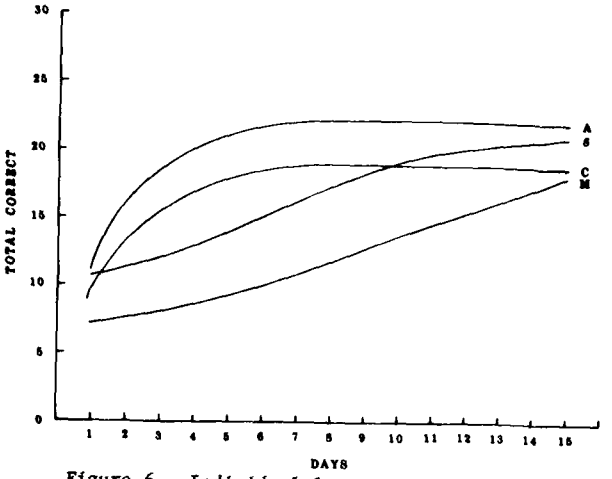


Figure 6. Individual learning curves of 4 subjects (8, A, C, M) who improve at different rates, but reach terminal levels correlated with initial performance over 15 days on the Grammatical Reasoning Test.

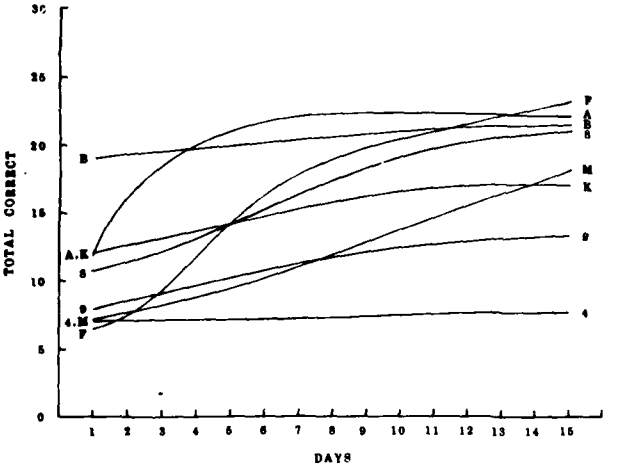


Figure 7. Individual learning curves showing differences in initial and terminal levels, and rates of learning for 8 subjects (4, 8, 9, A, B, F, K, M) over 15 days on the Grammatical Reasoning Test.

PROCEEDINGS OF THE 22ND ANNUAL MEETING OF THE HUMAN FACTORS SOCIETY
DETROIT, MI, OCTOBER, 1978

PROGRESS IN THE ANALYSIS OF A PERFORMANCE EVALUATION TEST FOR ENVIRONMENTAL RESEARCH (PETER)

Robert S. Kennedy and Alvah C. Bittner, Jr.
Naval Biodynamics Laboratory, New Orleans, LA 70189

INTRODUCTION

This report deals with the progress in the development of a Performance Evaluation Test for Environmental Research (PETER), a program motivated by the need for a test battery which is suitable for administration through extensive repetitions (Kennedy & Bittner, 1977). Nearly all studies into unusual environments employ subjects-as-their-own-control to the extent that "Environmental Time-Course" (ETC) effects may be considered paradigmatic of a class of studies which incorporates "repeatability" as a characteristic ingredient. Stated differently, with these paradigms the concern is chiefly with the effects of an environment on performance. The effect of exposure duration itself is nearly always included as an unwanted consequence of the experiment. Although much research has been conducted using an ETC paradigm, most of it had been accomplished with batteries insufficiently standardized to yield unambiguous results. Additionally, related literature concerning time course changes in skill acquisition (cf. Jones, 1962, 1969, for example) could profitably be incorporated into those studies which follow an ETC paradigm.

Standardization of PETER is being accomplished to provide intercorrelation reliabilities obtained over 15 days of testing, in addition to means and standard deviations (Kennedy & Bittner, 1978). The tests which have been selected for study early in our program sample cognitive, perceptual and information processing functions. Psychomotor, sensory and physical proficiency tasks will be studied later. The purposes of the present paper are: (1) to describe our experiences with the first ten tasks we have studied; and (2) to make inferences about the implications of these results for environmental research in general.

METHOD

A cadre of 19 Navy enlisted men, ages 19 to 24, were tested for 15 consecutive weekdays. Tests on one, or at most two of the ten tasks were administered each day, with testing performed in the morning between 8 a.m. and 10 a.m. Subjects were monitored for fitness by a team of physicians. All volunteer subjects were recruited and evaluated in accordance with procedures specified in Secretary of the Navy Instruction 3900.39 and Bureau of Medicine and Surgery Intruction 3900.6 which require voluntary informed consent and meet or exceed the most stringent provisions of all prevailing national and international guidelines.

RESULTS

Complex Counting Test (Kennedy & Bruns, 1975)

Results for this test are shown in Figures 1 and 2. Both mean scores and standard deviations (Figure 1) were relatively level (within 10 percent) over the three weeks of testing. Correlations are shown in Figure 2, where performances on selected base days (Days 1, 2, 4, 9, and 13) are compared with each subsequent day, not only in order to determine intertrial reliability of a particular day's performance, but also to monitor series effects in these reliabilities. Examining Figure 2, it may be seen that the reliability of Day 4 with subsequent days is very good ($r = >.85$). In Table I, the ANOVA shows a significant subjects effect ($p < 10^{-5}$) but a nonsignificant ($p > .10$) days effect.

Table 1

ANOVA: Complex Counting

SOURCE	DF	MS	F	P
DAYS	14	30.23	0.58	NS
SUBJS	18	2510.51	48.51	$<10^{-5}$
RESID	252	51.75		

Grammatical Reasoning Test (Baddeley, 1968; Rose, 1974)

The results are shown in Figures 3 and 4. Examining Figure 3, it may be seen that both means and standard deviations of performance increase over trials at a declining rate. The ANOVA in Table 2 supports this learning curve with a significant days effect ($p < 10^{-5}$), and also shows a significant subjects effect ($p < 10^{-5}$). Reliabilities are shown in Figure 4 and are moderate when comparing Days 1 and 2 with other days, but very good ($r > .80$) with comparisons made after Day 4. However, all reliabilities decline over sessions, (cf Jones, 1969) and the rates of decline are nearly equivalent.

Table 2

ANOVA: Grammatical Reasoning

SOURCE	DF	MS	F	P
DAYS	14	99.20	14.70	$<10^{-5}$
SUBJS	17	277.88	41.18	$<10^{-5}$
RESID	238	6.75		

Research performed under Navy Work Unit No. MF58.524-002-5027. The opinions are those of the authors and do not necessarily reflect those of the Department of the Navy.

Code Substitution Test (after Weschler, 1955)

The results appear in Figures 5 and 6. Mean performance (Figure 5) (total correct) improves over the 15 testing administrations but appears to decelerate after Day 9. Standard deviations (Figure 5) appear equal after Day 7. Average correlations (Figure 6) for subsequent days are poorest for Days 1 and 2. The reliability of Day 4 with later days is about .60. Table 3 contains an ANOVA for total correct and shows significant days and subjects effects.

Table 3

ANOVA: Code Substitution

SOURCE	DF	MS	F	P
DAYS	14	504.09	7.71	$<10^{-5}$
SUBJS	18	1524.76	23.32	$<10^{-5}$
RESID	252	65.37		

Stroop Test (Jensen & Rohwer, 1966)

The data appear in Figures 7, 8, 9, and 10. Mean scores for three directly measured performances and two difference scores (derived) are shown in Figure 7. Performance improves for 10 days on all measures but appears relatively asymptotic thereafter. Standard deviation scores are found in Figure 8. The correlations were highest for colored blocks (CB) and poorest for the derived score CB-CW (Figure 10). Correlations for colored words (CW), the most commonly used *score*, are shown in Figure 9. The present test administration differed from that used by most other investigators in that response keys (vice verbal responses) were used and test administrations were brief (30 seconds), and may have been a factor in obtaining lower reliabilities than reported elsewhere (Jensen & Rohwer, 1966). The ANOVAs for all Stroop scores showed significant subjects and days effects and two (CW & CB - CW) are shown in Tables 4 and 5 respectively.

Table 4

ANOVA: Stroop Test Color Words

SOURCE	DF	MS	F	P
DAYS	14	657.64	29.11	$<10^{-5}$
SUBJS	18	1356.63	59.15	$<10^{-5}$
RESID	252	22.93		

Table 5

ANOVA: Stroop Test CB-CW Score

SOURCE	DF	MS	F	P
DAYS	14	173.77	5.43	$<10^{-5}$
SUBJS	18	198.71	6.20	$<10^{-5}$
RESID	252	32.03		

Arithmetic Test

This was a paper and pencil test which alternated arithmetic operations: three digit addition; three digit subtraction; two digit by two digit multiplication; and four digit by two digit division. Figure 11 shows mean performances which appeared to be reaching an asymptote after 10 days of testing. The standard deviations also shown in Figure 11 appear to increase throughout the experiment suggesting that dispersion increases over sessions. The reliabilities (Figure 12) are generally high ($r > .90$) and do not appear to decline over sessions. Table 6 shows days and subjects effects. It is of interest that both number attempted, number correct and number right minus wrong, reflected average reliabilities substantially higher ($r > .90$) than percent correct of number attempted ($r < .70$) for the same data.

Table 6

ANOVA: Arithmetic Test

SOURCE	DF	MS	F	P
DAYS	14	233.88	8.17	$<10^{-5}$
SUBJS	17	4850.05	169.35	$<10^{-5}$
RESID	238	28.64		

Neisser Letter Search (Neisser, Novick & Lazar, 1963; Rose, 1974)

Results are shown in Figures 13 and 14. Mean slope scores and standard deviations shown in Figure 13 appear relatively level for the duration of the experiment although with some variability. Means and standard deviations also seem to co-vary. In Figure 14, correlations were low for base days 1, 2 & 4 ($r = <.50$) but appeared higher after Day 9. Table 7 shows significant subjects and days effects.

Table 7

ANOVA: Letter Search

SOURCE	DF	MS	F	P
DAYS	14	.15	8.93	$<10^{-5}$
SUBJS	17	.13	8.07	$<10^{-5}$
RESID	238	.02		

Critical Tracking Test (Jex, McDonnell & Phatak, 1966; Rose, 1974)

The data appear in Figures 15 and 16. Mean scores (Figure 15) improve for the duration of the experiment but at a declining rate. The plateau on Days 13 through 15 is due either to performance reaching an asymptotic level or to the subjects anticipation of the completion of the experiment. The standard deviation (Figure 15) was relatively constant over days. The average reliability (Figure 16) of Days 1 and 2 with subsequent days is far lower ($r < .60$) than for Day 4 and thereafter. The decline over days is very apparent. The ANOVA (Table 8) shows significant days and subjects effects.

Table 8

ANOVA: Critical Tracking

SOURCE	DF	MS	F	P
DAYS	14	9.74	49.87	$< 10^{-5}$
SUBJS	17	6.79	34.76	$< 10^{-5}$
RESID	238	.20		

Subcritical Two Dimensional Compensatory Tracking Test

This test was administered after the completion of the critical tracking test. An acceleration control displacement stick was used. Mean and standard deviation scores reached a plateau (Figure 17) by Day 5. Reliabilities (Figure 18) were high the first 10 days, but apparatus malfunction produced a dead spot on the CRT which was discovered by a few subjects around Day 10. Thereafter reliabilities degraded. Both subjects and days effects were significant in the ANOVA (Table 9).

Table 9

ANOVA: Compensatory Tracking Test

SOURCE	DF	MS	F	P
DAYS	14	57.60	42.69	$< 10^{-5}$
SUBJS	17	10.41	7.72	$< 10^{-5}$
RESID	238	1.35		

Time Estimation (Graybiel, et al., 1965)

Results are shown in Figures 19 and 20. The means and standard deviations (Figure 19) were relatively level. Table 10, summarizes the ANOVA and indicates that the subjects effect was significant, however, the days effect was not significant. These results support data obtained previously (Graybiel, et al., 1965).

Table 10

ANOVA: Time Estimation est

SOURCE	DF	MS	F	P
DAYS	14	.88	.85	NS
SUBJS	18	7.51	7.30	$< 10^{-5}$
RESID	252	1.03		

Reliabilities of given base days with each subsequent day were moderate but approached zero with additional days (Figure 20). A fine grained analysis (McCauley, Kennedy, & Bittner, in press) shows that parts of this test have higher reliabilities ($r > .90$) than the whole test.

Spoke Test

This test is a modification of the Trail Making Test (Reitan, 1955) and has a psychomotor subtask, the control task (CT), and a visual search subtask, the experimental task (ET). Figure 21 shows level mean scores and slight variability in standard deviations for the CT measure. However, the days effect, in addition to the subjects effect, was significant (Table 11). Figure 22 shows level and moderately high reliabilities which do not appear to increase or decrease with trials after Day 2.

Table 11

ANOVA: Spoke Test Control Task

SOURCE	DF	MS	F	P
DAYS	14	28.09	3.13	$< 10^{-4}$
SUBJS	17	399.84	44.62	$< 10^{-5}$
RESID	238	8.96		

Figure 23 shows improving search times over the first few days and relatively level performance thereafter. The ET standard deviations were somewhat variable. Table 12 shows significant days and subjects effects for ET. Reliabilities for ET (Figure 24) were lower than CT ($r < .30$) for Base Day 4 and thereafter.

Table 12

ANOVA: Spoke Test Experimental Task

SOURCE	DF	MS	F	P
DAYS	14	1388.22	5.24	$< 10^{-5}$
SUBJS	17	2938.26	11.10	$< 10^{-5}$
RESID	238	264.68		

DISCUSSION

Fifteen measures on ten different tests were reported in this study, thirteen of which showed significant learning (i.e., days) effects. The two exceptions were Time Estimation and Complex Counting, replicating findings reported elsewhere (Kennedy & Bruns, 1974; Graybiel, et al., 1965). The greatest practice effects appeared with both tracking tests, followed by the Stroop Test and the Grammatical Reasoning Test, tasks on which reaction time or speed of manual response could contribute to the total score. An inspection of the variations in obtained standard deviations over sessions exemplifies the importance of testing control groups. Some standard deviations remained level after a few days; some co-varied with other measures of performance; and other showed no systematic trends related to changes in the means or to changes in the reliabilities of the tests. The analyses of reliabilities over extensive testing showed that they were sufficiently high for the inclusion of some tests in a battery in their present form (e.g., Complex Counting and Arithmetic) and suggest that longer tests may be required for others (Coding and Spoke ET). In some cases, reliabilities degrade to a point (Time Estimation, Stroop derived scores) that it is unlikely that an effect however large, could be shown to be statistically significant if the test were employed in its present form. Previous environmental research with test batteries can be questioned based upon the results shown in this report. The decline in reliabilities of tasks with repeated testing shown for most tasks in this study, indicates that the "factors" measured in an experiment may change over time. Control groups provide protection against changes in mean performances, but the responses of subjects in both experimental and control groups may reflect one "factor" at the beginning (X) and another at the end (Y). Differences, therefore, may be due to mean differences in (X) initially and in (Y) at the end. High reliability over only one test repetition affords little protection from this problem. Time Estimation, for example, showed high reliability (r=.95) for the relationship of Base Days 9 and 10 (i.e., after eight days practice). However, the full regression (to r = .60) with only six administrations) showed that only by long term studies, such as the present one, can experimental tasks be evaluated for meaningful application in environmental research.

REFERENCES

Baddeley, A. D. A 3 minute reasoning test based on grammatical transformation. Psychonomic Science, 1968, 10, 341-342.

Graybiel, A., Kennedy, R. S., Knoblock, E. C., Guedry, F. E., Jr., Mertz, W., McLeod, M. E., Colehour, J. K., Miller, E. F., III, & Fregly, A. R. Effects of exposure to a rotating environment (10 rpm) on four aviators for a period of twelve days. Aerospace Medicine, 1965, 36, 733-754.

Jensen, A. R., & Rohwer, W. D. The Stroop Color-Word Test: A review. Acta Psychologica, 1966, 25, 36-93.

Jex, J. R., McDonnell, J. D., & Phatak, A. V. A "critical" tracking task for manual control research. IEEE Transactions on Human Factors in Electronics, 1966, HFE-7, 138-145.

Jones, M. B. Practice as a process of simplification. Psychological Review, 1962, 69, 274-294.

Jones, M. B. Differential processes in acquisition. In E. A. Bilodeau and I. McD. Bilodeau (Eds.), Principles of Skill Acquisition. New York: Academic Press, 1969.

Kennedy, R. S., & Bittner, A. C., Jr. The stability of complex human performance for extended periods: Application for studies of environmental stress. Proceedings of the 49th Scientific Meeting of the Aerospace Medical Association, New Orleans, LA, May 1978. Washington, D.C.: Aerospace Medical Association.

Kennedy, R. S., & Bittner, A. C., Jr. The development of a Navy Performance Evaluation Test for Environmental Research (PETER), In Pope, L. T., & Meister, D. (Eds.), Productivity Enhancement: Personnel Performance Assessment in Navy Systems, Naval Personnel R & D Center, San Diego, CA, 12-14 October 1977. (NTIS No. AD A056047)

Kennedy, R. S., & Bruns, R. A. Some practical considerations for performance tested in exotic envronments. In B. O. Hartman (Ed.), Higher Mental Functioning in Operational Environments. AGARD Conference Proceedings No. 181, October 1975. (Advisory Group for Aerospace Research and Development, Organization du Taite de L'Atlantique Nord).

McCauley, M. E., Kennedy, R. S., & Bittner,, A. C., Jr. Development of a Performance Evaluation Test for Environmental Research (PETER): Time Estimation Test. Perceptual and Motor Skills (in press).

Neisser, U., Novick, R., & Lazar, R. Searching for ten targets simultaneously. Perceptual and Motor Skills, 1963, 17, 955-961.

Reitan, R. M. An investigation of the validity of Halstead's measures of biological intelligence. Archives of Neurology and Psychiatry, 1955, 73, 28-35.

Rose, A. M. Human information processing: An assessment and research battery. Ann Arbor, MI, University of Michigan, Doctoral dissertation, 1974, (also published as AFOSR-PR-74-1372). (NTIS No. AD785411)

Wechsler, D. Manual for the Wechsler Adult Intelligence Scale. New York: Psychological Corporation, 1955.

FIGURES

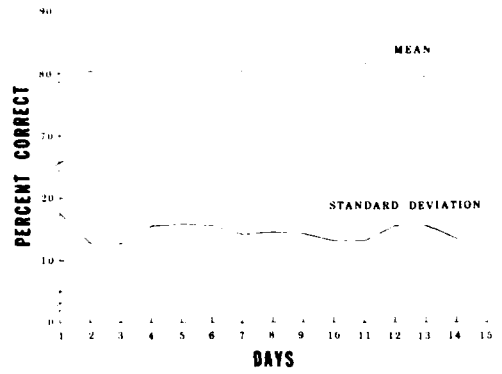


Figure 1. Complex Counting Test means and standard deviations for percent correct over 15 days (n=19).

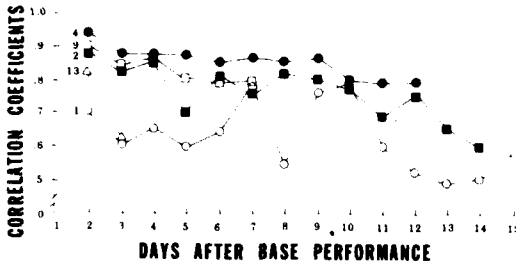


Figure 2. Complex Counting Test correlations for selected base days (1, 2, 4, 9, 13) and those following for percent correct over 15 days (n=19).

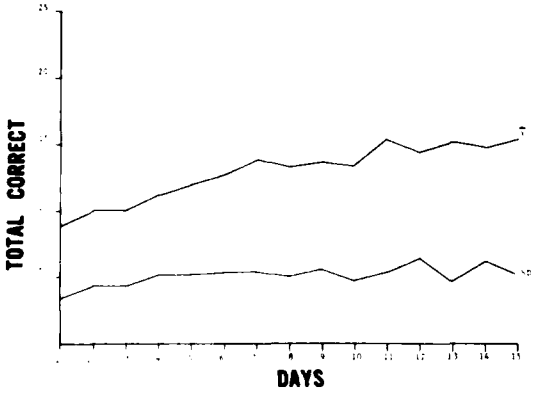


Figure 3. Grammatical Reasoning Test means and standard deviation for total correct over 15 days (n=18).

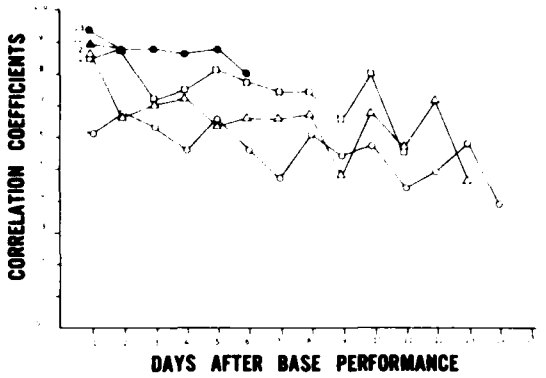


Figure 4. Grammatical Reasoning Test correlations for selected base days (1, 2, 4, 9, 13) and those following for total correct over 15 days (n=18).

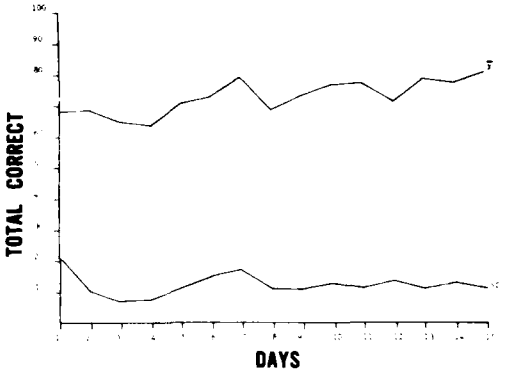


Figure 5. Code Substitution Test means and standard deviations for total correct over 15 days (n=18).

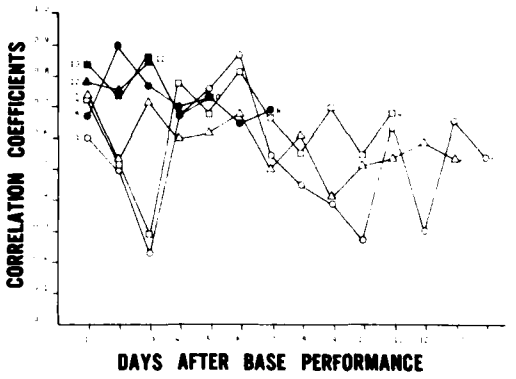


Figure 6. Code Substitution Test correlations for selected base days (1, 2, 4, 8, 10, 12) and those following for total correct over 15 days (n=18).

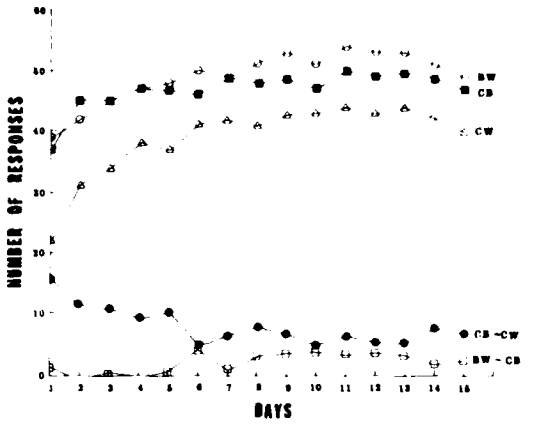


Figure 7. Stroop Test means for number of responses on five measures: black and white words (BW), color blocks (CB), color words (CW), CB-CW, and BW-CB, over 15 days (n=19).

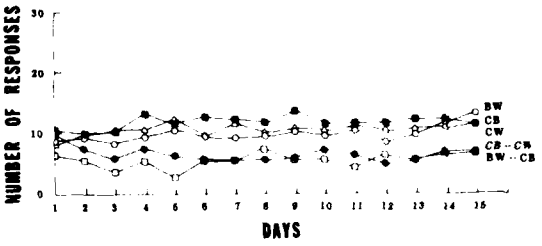


Figure 8. Stroop Test standard deviations for 5 measures over 15 days (n=19).

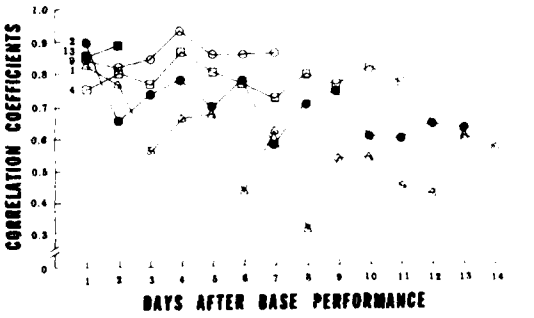


Figure 9. Stroop Test correlations for selected base days (1, 2, 4, 9, 13) and those following for colored words over 15 days (n=19).

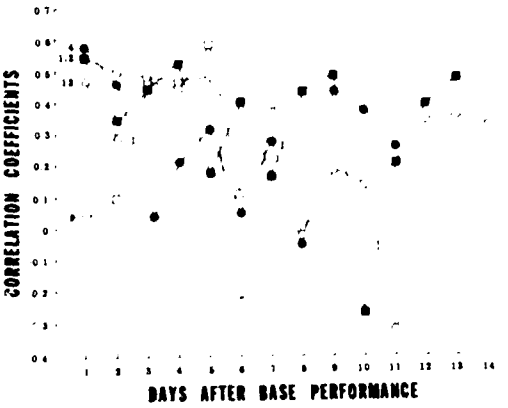


Figure 10. Stroop Test correlations for selected base days (1, 2, 4, 9, 13) and those following for CB-CW over 15 days (n=19).

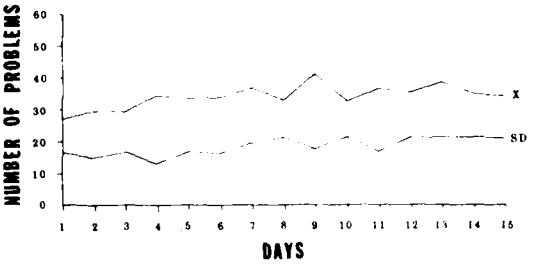


Figure 11. Arithmetic Test means and standard deviations for total correct over 15 days (n=18).

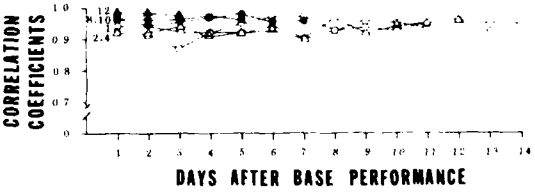


Figure 12. Arithmetic Test correlations for selected base days (1, 2, 4, 8, 10, 12) and those following for total correct over 15 days (n=18).

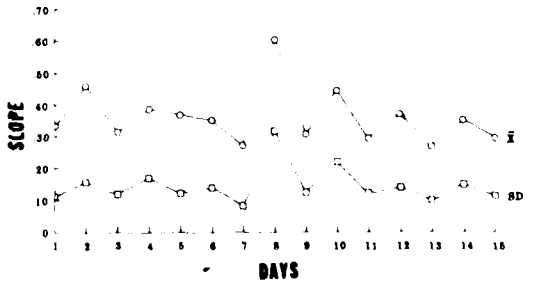


Figure 13. Letter Search Test means and standard deviations for time per item slope over 15 days (n=18).

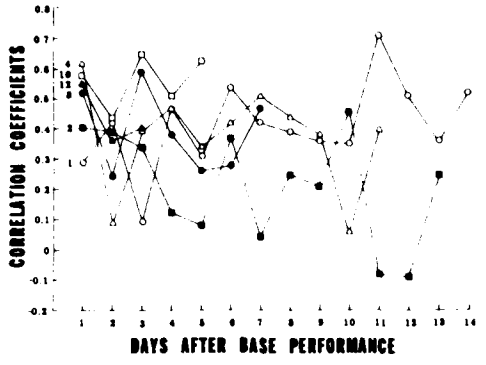


Figure 14. Letter Search Test correlations for selected base days (1, 2, 4, 8, 10, 12) and those following per time per item slope over 15 days (n=18).

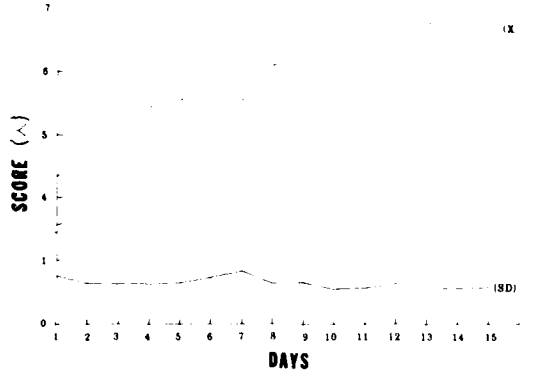


Figure 15. Critical Tracking Test means and standard deviations of scores over 15 days (n=18).

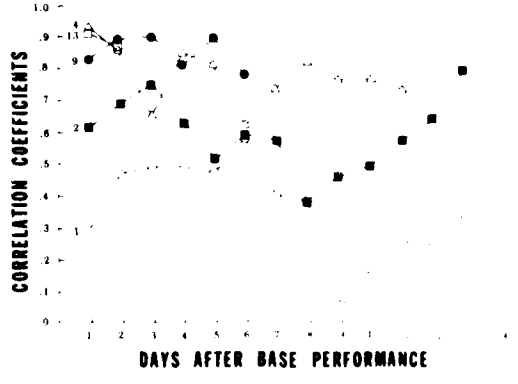


Figure 16. Critical Tracking Test correlations for selected base days (1, 2, 4, 9, 13) and those following for scores over 15 days (n=18).

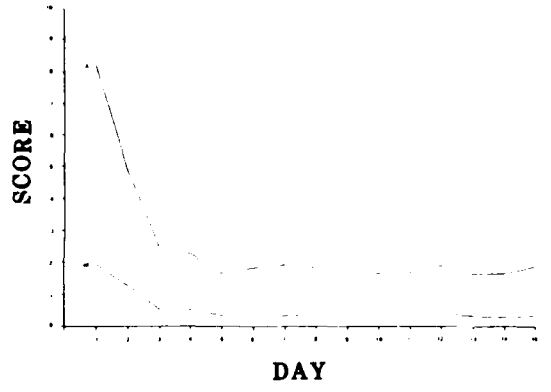


Figure 17. Compensatory Tracking Test means and standard deviations for RMS error over 15 days (n=18).

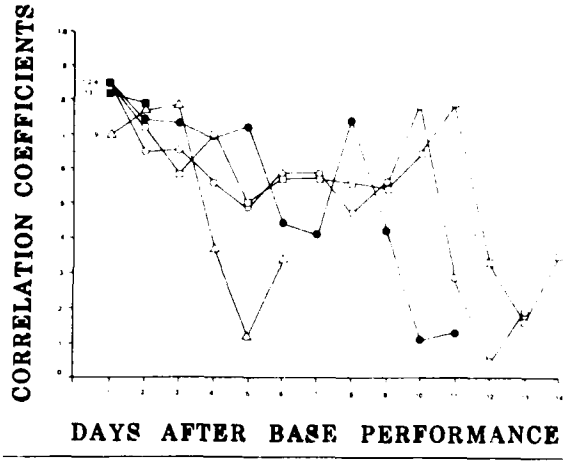


Figure 18. Compensatory Tracking Test correlations for selected base days (1, 2, 4, 9, 13) and those following for RMS error over 15 days (n=18).

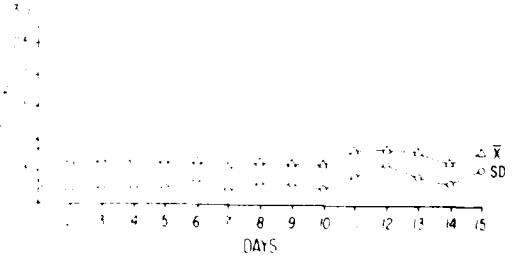


Figure 19. Time Estimation Test means and standard deviations for constant error over 15 days (n=19).

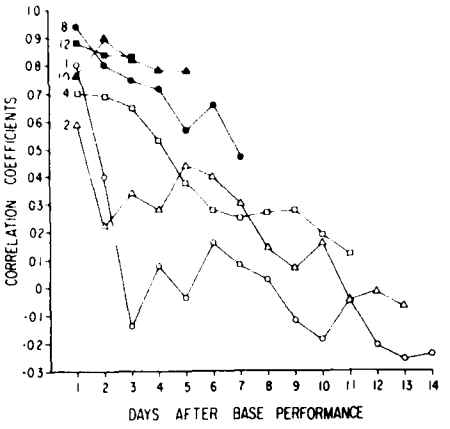


Figure 20. Time Estimation Test correlations for selected base days (1, 2, 4, 8, 10, 12) and those following for constant error over 15 days (n=19).

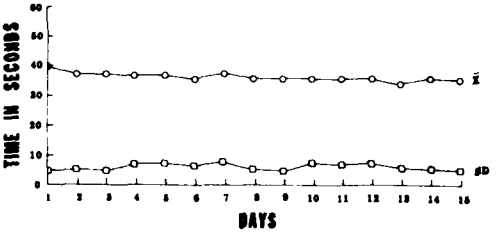


Figure 21. Spoke Test means and standard deviations for control task (time to completion) over 15 days (n=18).

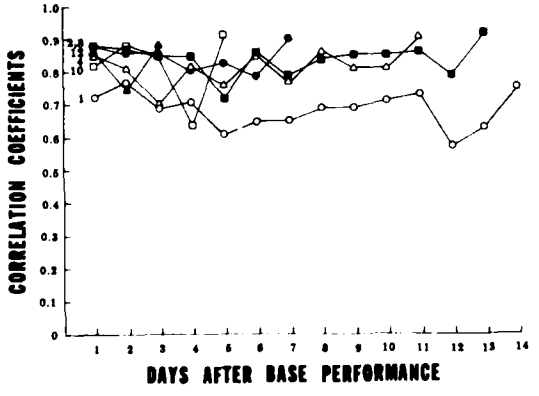


Figure 22. Spoke Test correlations for selected base days (, 1,2 4, 8, 10, 12) and those following for control task over 15 days (n=18).

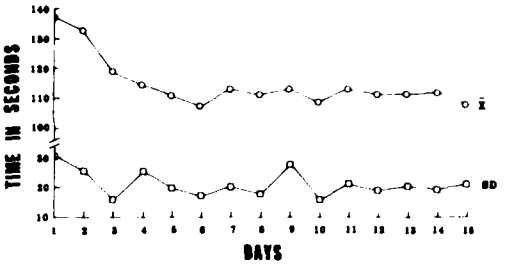


Figure 23. Spoke Test means and standard deviations for experimental task (time to completion) over 15 days (n=18).

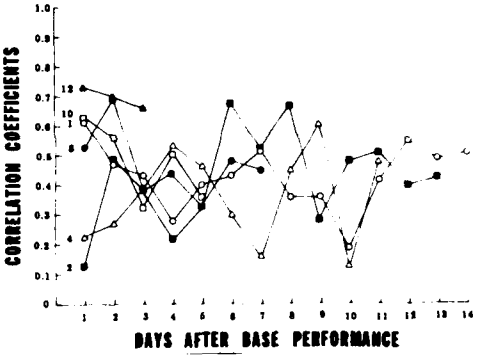


Figure 24. Spoke Test correlations for selected base days (1, 2, 4, 8, 10, 12) and those following for experimental task over 15 days (n=18).

Symposium Proceedings: Productivity Enhancement: Personnel Performance Assessment in Navy Systems
12 - 14 Oct 77 Navy Personnel Research & Development Center, San Diego, CA 92152

THE DEVELOPMENT OF A NAVY PERFORMANCE EVALUATION TEST FOR ENVIRONMENTAL RESEARCH (PETER)

Robert S. Kennedy, CDR MSC USN
Head Human Performance Division and Officer-in-Charge
Naval Aerospace Medical Research Laboratory Detachment
and
Alvah C. Bittner, Jr.
Human Factors Engineering Branch
Point Mugu Test Center

ABSTRACT

The basic problem with performance testing in exotic environments is the general unwillingness of investigators to take the time to standardize a test battery. Many other problems exist and are obvious to all who have tried to measure performance under usual and unusual environmental conditions. It is the purpose of this paper to set forth some of the problems that have grown out of our experiences and which we feel have not been extensively commented upon in the research literature, and also to describe our plan for solution.

Preface

The present plan is a simple one: The literature will be searched for human performance tasks which have been shown to degrade under motion (vibration and ship motion), during thermal exposure, and under pressure. The performances that meet these first criteria will be categorized as cognitive (decision making, information processing, judgment), motor (tracking, reaching), etc., and a taxonomy of performances will be developed. Additionally, each performance task will be evaluated in the following way: 20 subjects will be tested 10 times (5 days/week for 2 weeks) to determine three types of reliability: internal consistency, the accuracy and sensitivity to separate individuals, and the stability of this accuracy and sensitivity over repeated testing. Performances on these tasks will be compared to scores on other tests of mental functions. Progress to date will be reported.

The National Aeronautics and Space Administration, the Advanced Research Project Agency, the Navy (via the Office of Naval Research), and the Bureau of Medicine and Surgery have funded several studies (see Kennedy, 1977 for a review) which have nearly all made very similar points regarding the standardization of a performance test battery for assessment of environmental stressors. In the main, test batteries have been proposed, particularly factor analyzed batteries, but rarely have normative data been collected and never have practice effects been studied effectively.

The original title for the present paper was very broad and included all Navy R & D concerning performance. We intend, however, merely to present how the Naval Aerospace Medical Research Laboratory Detachment plans to research the general area, with specific application to our interests in the effects of ship motion or performance. It should be noted that, in addition to the human performance R & D already presented at this symposium by various members of the Navy

Personnel Research and Development Center, complementary programs also exist within the Engineering Psychology Programs of the Office of Naval Research and within the Human Effectiveness Programs of the Naval Medical Research and Development Command.

INTRODUCTION

Casual observation over several years of performance testing and a comprehensive reading of over 400 "human performance studies" in hyperbaria (see Bachrach & Kennedy, 1977, for a review) suggest that there is a need for future studies into the standardization of a human performance test battery.

In our opinion, the persons who initiated the experiments requiring performance testing in exotic environments were generally persons who became involved originally because of a primary interest in the environment rather than in the performance. (Within "environment" we include unusual sensory stimulations, drugs, fatigue, and even learning, as well as motion sickness, hyperbaric, etc.) Thus, we feel that, frequently, several criteria were employed (often trading back and forth among them) in the selection of tasks for inclusion in a battery to be assembled. These criteria have included the following:

1. Literature findings that were recollected, probably because the results of tests were unusual.

2. What colleagues and friends had done.

3. What demonstration experiments were performed in experimental psychology laboratory during their student days.

4. Chapter headings in Woodworth and Schlosberg (1954) and other standard texts.

5. Equipment left behind in the storage room of the laboratory by their predecessors.

6. That which could be quickly and easily assembled from clever ideas, (the so-called toy gadget approach).

7. Stock items from apparatus companies.

8. Logistic limitations forced by the environment or project (e.g., small, inexpensive, no tubes, portable, nonmagnetic, self-scored, no sparks, self-administered, battery powered, and rugged).

9. Similar to the work done by real-world persons.

10. A relatively basic kind of skill is involved; that is, learning theoretically SHOULD be able to be accomplished quickly.

11. Less often, performances could be expected to be disrupted on the task in this environment.

We believe that the criteria listed above have been employed often enough to assemble batteries so that these criteria are worth citing. It should also be noted, however, that, typically, a test battery was generally an ad hoc response to the imminent availability of an environmental condition, whether the environment was a hurricane (Kennedy, Moroney, Bale, Gregoire, & Smith, 1970), a rotating room (Guedry, Kennedy, Harris, & Graybiel, 1964; Fregly & Kennedy, 1965; Kennedy, Tolhurst & Graybiel, 1965), or a deep dive. Thus, long-range planning frequently is not possible. In summary, it is felt that performance test batteries are often assembled for largely practical reasons, on short notice, by persons whose major interest is not performance testing. To alleviate these problems we have combined, in tabular form, what we consider the traditional, important criteria for test construction along with the practical aspects concerning operational performance assessment. These criteria are summarized in Tables 1 - 4. In addition, other problems with performance test battery construction exist.

1. What performance tests are designed to measure

Although this distinction is not generally made, it is implicit that performance testing is undertaken for two main purposes: first, to be able to make some statement about the integrity of the organism, and second, to determine whether an environment interacts with an organism's ability to do a particular kind of work (cf. Table 3). In this paper, the first purpose will be called "CNS status," and the second, "effectiveness of a system's output." Examples of tests designed for the former purpose include reaction time, digit span, tremor, electroencephalogram, speed of tapping, and CFF. Examples of the latter include an underwater pipe puzzle, a sonar monitoring task, Morse code tests, and speech intelligibility tasks. Frequently, both types of tasks are included in a single experiment into the environment's effect on man and without regard to the distinction made above. The advantage of the latter approach is that the system's concept is used and the translation to real-activities is direct. (Also, subject cooperation is usually better.) The disadvantage is that no general principles are adduced and the application of the findings holds only for the stimulus condition employed. For instance, tracking studies with CRT displays have been conducted for many years and very few general rules have resulted (Adams, 1961). The major disadvantage of the first approach (index of an organism's integrity) is that they depend heavily upon the knowledge of the validity of the task. If only face validity is available, other considerations (money, size, apparatus, and availability) must be used to justify inclusion. If face validity is not evident, then justification is very tenuous.

The distinction made between these two strategies is subtle, but it is also real, and its existence complicates the results of many studies. This is chiefly due to the fact that the two approaches require different research philosophies, although the ultimate aim of both approaches is similar: namely, prediction (i.e., an ability to account for 100 percent of the variance).

The first approach comes directly from experimental psychology and usually follows an analysis of variance model. Thus, the numerous tests in a test battery are designed to sample all of the skills (factors) of the organism. The implication is that, if the full range of human abilities is tested, one can generalize the findings and apply them to other circumstances (e.g., subjects, treatments, etc.). This approach depends heavily upon following the principles of test

construction: (1) norms, (2) reliabilities, (3) validities, (4) factors tested, (5) effects of practice, and (6) individual differences. If all these principles were satisfactorily fulfilled, it would be possible to employ the test in an exotic environment and account for all the main effects of such an environment on human performance. For example, if it were known that hand dynamometry correlated perfectly with all other kinds of voluntary skeletal muscle output, and the Harvard Step Test (Kennedy & Hutchins, 1971) with all cardiac muscle output, then it would not be necessary to use other tests of these functions. The difficulty, of course, is that neither of these tests correlates sufficiently. Additionally, other "more psychomotor" tasks are even less clear-cut with regard to what they are measuring (i.e., validities). However, the problem does not end here. Reliabilities of a test battery--any test battery--are not completely known. No norms (expected values) are available on a sizable population, particularly when practice effects are concerned. However, factor analyses studies (e.g., those of Fleischman) have been completed for some samples.[1]

The second approach is in vogue more now than previously, probably because it emphasizes a systems approach. The statistical model employed is correlation, and in general, single factor studies are conducted. The overall plan is to replicate real-world work and to do it under controlled conditions. The second approach does not depend upon the validity of the task as heavily as the first method, since it, itself, is the work. However, the characteristics of the subjects are critical. It is important, and usually essential, that the subjects be the same kind of people as the real-world workers toward whom the data will be applied. The shortcoming of this strategy is also its chief advantage: the application of the findings from such studies is specific and immediate, but sometimes it is so specific that generalization within the same environment, but with slight differences, may not be possible.

2. Two experimental paradigms

There are two main ways in which to study the effects of the environment on a subject's ability to do work. The first (most often used) uses the subject as his own control and generally follows a pre-, per- and post- paradigm. In the pretest, the subject is practiced on all the tests to be employed in order to arrive at a learning plateau. Then he is placed in the experimental situation to see whether or not it disrupts performance. Posttesting is used to monitor recovery effects, if there are any. There are many problems with this approach. Chiefly, psychomotor performance almost never arrives at a plateau. This is discussed in more detail later in this paper. Asymptotes occasionally are obtained, but these, too, are infrequent. Even on tests where one would expect practice to be accomplished quickly (e.g., reaction time, CFF, tracking visual acuity),[2] the environment itself occasionally causes certain tests to be performed less well while standing during rotation, and is probably also measuring

---

[1] Sinbad (1969) is based on these studies and, when standardized, may be used to obviate some of the problems mentioned above.

[2] The use of signal detection theory (Swet, Tanner, & Birdsall, 1961) as a methodology may be helpful here, but as we all know from the way the 100-yard dash record is continually broken, it is not just a criterion problem. Stated differently, a knowledge of sensory sensitivity, d' (d-prime) separated from the subject's criterion (beta) would refine present knowledge, but d', even carefully and prudently measured, may change with practice.

body sway (Graybiel, Kennedy, Knoblock, Guedry, Mertz, McLeod, Colehour, Miller, & Fregly, 1965). This point will also be discussed later. Post-effects also present difficulties since motivation changes (e.g., end spurt in vigilance) usually attend the imminent completion of an experiment.

The alternative approach: to test "just before" and "just after" the environmental exposure (say a 12-hour overwater ASW flight) has its own problems; namely, the experimenter feels that it is necessary to be aware of the status of the subject during the exposure. If the testing is short (e.g., hand dynamometry), it can be influenced by the bias of a subject and summoning efforts for a "one-shot-deal" so that, often, changes are not obtained even though the subject is frankly tired. If the testing period is long (e.g., treadmill), it can contribute to the fatigue. In addition, lengthy posttests are often unfair to the subject.

3. Assessment of input-integrator-output circuits

The general form of psychological experimentation follows an S-R paradigm, or SOR, where O is for organism (Graham, 1951). Performance testing employs this paradigm particularly when "CNS status" type experiments are conducted. Typically, in these studies the experimenter is mainly interested in whether his treatment (drugs, hypoxia, confinement, magnetic fields) produces any CNS change. So, a stimulus is presented and the output of the organism is monitored for changes. Frequently, however, due account is not taken as to whether the stimulus was adequately received by the receptor (retina, ear, hair cells, etc.) then properly delivered along that nerve pathway; also, whether the output (muscle) pathway is similarly unaffected. For example, during acceleration stress, the lack of oxygen to the retina indicates that signals are not adequately received at the receptor site. This also occurs with the differences obtained in visual performance underwater. The physical conduction of light in air versus water may account for these differences -- most likely the visual signal is just not delivered to the receptor in water as well as in air, so one would not posit CNS changes underwater to account for the poorer visual acuity obtained. At the other end of the nerve-muscle circuit, changes in four-choice reaction time done underwater clearly have the friction of water on the one hand to slow down performance as well as the possible other effects of compression and mixed gases and so, probably, CNS changes cannot adequately be assessed with this task. So, too, past pointing underwater may be different: not because of central involvement, but because of inertial differences on the arm. This is not to imply that such studies should not be undertaken, rather, it behooves the experimenter to indicate where possible which part of the OSR circuit he is testing. Therefore, one must know about the transmission characteristics of light, the dependency of the retina on oxygen, and the viscosity and buoyancy characteristics of water. However, if such tasks are included in batteries that have other tests, (the intention of which is to tap the state of the CNS) when all results are reported together, there is confusion.

It would be useful to other investigators if results of experiments were reported relative to that part of the circuit which is being tested. This cannot be done in all cases, but it is possible to improve present reporting practices. Perhaps if we intellectually remove the known physical environmental effects from the periphery (nerve and muscle), we may be left with the finding that motivation

and the partial pressure of oxygen in the brain are the chief contributors to performance decrement under all conditions. The above criticism does not apply to the "systems output" type of studies which take no position regarding where in the circuit the problem occurs. Rather, their sole purpose is to determine whether an interaction of environmental condition occurs on people doing work. It is proposed that "CNS status" be used as a term to be contracted with "input/output quality" types of studies, whereby the former would deal with throughput changes due to the environment and the latter would address the physical aspects of the environment on man.

4. Practice effects

In a significant but not widely referenced paper, Bradley (1962) reported the persistence of sequence effects during psychomotor testing. Virtually all who study performance over many sessions have obtained similar findings. As was mentioned earlier, the investigator usually performs baseline pretesting before placing the subjects in the environment. Often, many trials are given (in one study, 7 days of testing) in an effort to have performance asymptotic "so that the pimple on the line can be more easily seen."[3] What is usually obtained is the well-known learning curve, which may, but does not always, asymptote. The problem with this approach is obvious, but there is another less obvious problem; that is, performance on a task after many trials is probably no longer an index of the same activity or place in the CNS that it was initially.

Studies by Ades and Raab, 1949, on the Kluver Bucy Syndrome (cited in Bachrach and Kennedy, 1977) illustrate the latter point where animals with certain portions of their brains removed were able to perform a visual discrimination task about as well as unoperated animals; however a simularly operated group was never able to learn this task.

Moreover, it is well known from the learning literature that, with extended practice, subjects overlearn, and when something is overlearned, it becomes more resistant to extinction. Therefore, for performance testing in exotic environments, if intensive practice is given on the tests prior to their use in the experimental environment, two factors appear inevitable: (1) the work is not an index of what it was at first, and (2) disruption of performance becomes very difficult. An example of this is as follows: move the index (first) and ring (third) fingers preferred hand together with the palms resting on a flat surface. Then move the second and fourth fingers together. Then, alternate 1 and 3, then 2 and 4, etc. Everyone can do this work, but it requires far more concentration for the average person than for a person who frequently plays the piano. The investigators believe that control for this activity is exerted high in the cortex for nonpianists, but has perhaps been shunted to a lower center in the CNS in practiced pianists. If the above is similar to what occurs in performance testing studies, the implications are obvious.

Because of the problems listed above, the following approach is planned: We feel that the approach is innovative, but it will draw heavily on the research literature for the initial selection of tests to be included for further study.

---

[3]Radloff, 1971, personal communication.

Those tests will be selected from the literature that meet criteria in one of the following areas: (1) demonstrated sensitivity to either thermal, motion, or hyperbaric environments by exhibiting degraded performances, (2) diagnostic capability (i.e., brain-damaged individuals have been found to perform differently from a normal population), and (3) measurement capability of a parameter of human information processing. After initial selection of the tests, the most promising will be subjected to further tests. The test and equipment attributes of each test will be viewed from the standpoint of the following factors ranked in general order of importance: (1) reliability (e.g., test-retest, alternate form, between and within administrations), (2) validity (e.g., predictive, context, construct, diagnostic-concurrent, fact), (3) other practical test factors (range of capability levels covered, sensitivity, transportability, efficiency), (4) equipment factors (e.g., availability, equipment reliability, transformability, safety, economy). Those tests that demonstrate a high level of adequacy on the above criteria will comprise an experimental battery. Performances on this battery will be compared to performances on a factor pure (e.g., Sinbad) battery to determine uniqueness of factors. Paper and pencil tests of cognitive functions (e.g., Bender-Gestalt, Guillford-Zimmerman) as well as well-standardized intelligence tests (e.g., Wais, Ravens, Stanford-Binet, Reitan, Halstead, Wunderlich) will be administered to this same population to further delineate and validate the factors obtained.

The first test that we have selected for further study is the so-called Beeper reviewed by Kennedy and Bruns (1975). The reasons for selecting this test originate partly from the literature review and partly from the study of acceleration stress by the NAS/NRC Committee on Bio-Astronautics, who convened a working group headed by Robert Galambos to discuss and report on principles and problems of performance testing. Using criteria based largely on earlier suggestions of Broadbent (1953), a performance test battery was proposed that would have general and specific applications.

We looked into Broadbent's report for ideas relative to the common problems of motion and acceleration stress and of exotic environments in general. Recommendations were also included for the use of tasks which are: "(a) work paced; (b) require vigilance; (c) over a long period of time; and (d) during which there is uncertainty in the stimulus display" (p. 22):

1. Laboratory norms on six different versions of this task for each of the approximately 100 college graduate males are available, as well as relationships to personality and other subject variables (e.g., hours of sleep) for these persons.

2. Neurophysiological correlates (vestibular nystagmus) of performance were shown.

3. Practice effects appear small on the three-channel auditory version and are known for the three-channel visual version.

4. The test can be group-administered.

5. It is relatively simple and inexpensive to construct.

6. There are many possibilities for constructing alternate forms.

7. Task difficulty can be controlled largely by instructions.

8. Latency of response within broad limits (namely, 1-2 seconds) is generally not a factor and so the task can appropriately be used even when environmental variables can interact physically with response speed (e.g., underwater).

9. Stimulus recording is binary and therefore is mechanically simple. Further, the regularity of the stimuli makes a scoring relatively easy and relatively independent of where on the magnetic tape a session begins.

10. Proportion measures are essentially linear (R .95) with absolute measures (namely, hits) and, therefore, direct comparisons can be made over different tasks.

11. Unlike many other vigilance tasks, many signals and responses occur and so individual time-line analyses are possible.

12. The results suggest that performance on forms of this task may be age-related.

The approach we have utilized includes the daily administration (15 minutes) of the Beeper for 2 weeks to study the reliability of the test in three ways: internal consistency, the accuracy and sensitivity to separate individuals, and stability of this accuracy and sensitivity over repeated testings.

We feel that this approach will serve as a model for future tasks to be included in our battery. At this writing, data are being collected, however the study is not completed. These results should be available at the meeting in October.

Table 1

Equipment Factors

Factors	Definition	References	Comments
Availability	Equipment software and hardware for presenting tasks, receiving responses, recording, scoring and integrating should be acquirable without excessive delays.	Alluisi (1967, 1969); Reilley & Cameron (1968); Kennedy (1971); Theologus et al. (1973)	Rose (1974) has suggested paradigm "reproducibility" (frequency with which a task has been studied) as a criteria for selection. Certainly selection of tasks most readily available in psychological laboratories would insure maximum cross laboratory availability. Some paper-and-pencil tasks rate high on this factor.
Equipment Reliability	Equipment software and hardware must be sufficiently reliable to permit sustained use for lengthy durations, i.e., have a high expected "mean time between failure." (MTBF)	Alluisi (1967, 1969); Theologus et al. (1973)	A method of checking hard and software states--proper or improper functioning--is a necessity for PTB tasks.
Transformability	Tasks can be adapted for administration in various environments of interest without seriously altering measurement capability.	Reilly & Cameron (1968)	Environments of interest could include "shirt sleeve laboratory," exotic environs (e.g., underwater), or field conditions. Portability (Rose, 1974) and potential for group administration (Kennedy, 1971) are valued elements.
Safety	Equipment should not present a potential health or safety hazard to subjects, and equipment must not be vulnerable to damage by stressed subjects.	Theologus et al. (1973)	This is the most important feature of any battery.
Economy	Costs for acquisition of equipment hard and software, administration, scoring, interpretation and maintenance should be reasonable.	Alluisi (1967, 1969); Reilley & Cameron (1968); Kennedy (1971); Theologus et al. (1973)	Temporal and monetary costs are important, albeit able to be traded off. Equipment based batteries have not been extensively applied or developed because of costs being excessive. Less expensive and sophisticated batteries would encourage standardizations of tasks in the literature.

Table 2

Reliability Factors

Factor	Definition	References	Comments
Test-Retest Reliability	Correlation established by administration of the same test to the same individuals on two different occasions.	Alluisi (1969) Grodsky (1967) Kennedy (1971) Theologus et al. (1973)	Experimenters using PTBs frequently administer the same task to subjects a large number of times. This has been shown in the literature to frequently result in changes in the nature of what is being measured and low correlations between early and later trials (cf., Woodrow, 1938 a & b; Fleishman and Hempel, 1955; Parker & Fleishman, 1960; Parker & Fleishman, 1961; Parker, 1964). Test-Retest reliabilities need to be determined over numbers of trials task will be administered.
Alternate-Form Reliability	Correlation established by administration of two "equivalent forms" of the same test (measuring same aspect of performance but of different questions or items) on two different occasions.	Theologus et al. (1973) Teichner (1974)	Alternate form reliability is appropriate for tasks with elements which lose potency when exposed to subjects. Also note that comment for Test-Retest applies with alternate forms which are employed over substantial numbers of trials.
Internal Consistency Reliability	Correlation estimate of the homogenity of a task's item scores established on a group of individuals on one occasion.	Theologus et al. (1973)	Note comment for Test-Retest has implication for internal consistency estimates a point delineated by Thorndike (1949).
Between Test Administrators Reliability	Correlation established by administration and scoring of the same or equivalent form of a task to the same individuals by two administrators.	Teichner (1974)	This reliability has not been of interest in most developments of PTBs, although the "experimenter effect" has a long history in experimental research (cf., Rosenthal, 1961).
Within Test Administrators Reliability	Correlation established by administration and scoring of the same or equivalent form of a task by the same administrator on two different occasions.	Teichner (1974)	This is the special case under which test Test-Retest and alternate forms reliabilities are established.

Table 3

Validity Factors

Factor	Definition	References	Comments
Predictive Validity	Correlation between operator performance on a task (or tasks) and future criterion performance or status.	Alluisi (1967, 1969) Grodsky (1967) Theologus et al. (1973)	"Real world" performance is a concern of experimenters who optimize this criteria vs. diagnosis of performance which is concerned with concurrent diagnostic validity.
Concurrent (Diagnostic) Validity	Correlation between test score and a diagnostic criterion status obtained at approximately the same time.	Teichner (1974)	Teichner (1976) stresses diagnostic aspects or tasks for assessment of subjects internal status (e.g., Is a nervous system dysfunction present?)
Content Validity	Extent to which a task or task battery covers a representative sample of the behavior domains to be measured.	Alluisi (1967, 1969) Reilley & Cameron (1968)	Related to content validity are the concepts of battery "generality" or "comprehensiveness" given as criteria by Theologus et al. (1973), Rose (1974) and Teichner (1974). These concepts stress that a battery should encompass as many critical aspects as possible while minimizing redundancy.
Construct Validity	Extent to which a test may be said to measure a "theoretical construct" or trait where theoretical construct or trait is established by convergence of information from a variety of sources.	Theologus et al. (1973) Rose (1974)	Rose (1974) has particularly stressed this concept by his emphasis on well used "paradigms" from experimental psychology with correlational and factor analysis as methods of convergence.
Factor Validity	Extent to which factor analysis indicates task as identifying or correlating with a factor.	Reilly & Cameron (1968) Theologus et al. (1973) Rose (1974)	Factor analysis has additional use in the assessment of the amount of redundancy in a battery.

Table 3 (Cont.)

Validity Factors

Factor	Definition	References	Comments
Face Validity	Extent to which test "looks valid" to subjects who take it, experimenters, or other observers	Alluisi (1967, 1969) Grodsky (1967) Reilly & Cameron (1968) Theologus et al. (1973)	Alluisi (1967, 1969) and Grodsky (1967) both stress need of face validity to insure subjects feel tasks are relevant and are motivated. *Theologus et al.* (1973), however, stresses the need of "...face validity to permit subjective generalization of effects...to the effects... on a 'real world' task..." Attempts to measure task face validity have not been reported. Briefing on importance of tasks vs. "face validity" method of motivating subjects have not been reported in literature although used as research strategy (e.g., by Cross & Bittner, 1969).

Table 4

Ability Range, Sensitivity, Trainability and Efficiency Factors

Factor	Definition	References	Comments
Range of Ability Levels Covered	Extent to which differing subject populations (varying in background, developmental level, training, etc.) can be tested.	Alluisi (1967, 1969) Reilly & Cameron (1968) Teichner (1974)	Although pointed out as important, this factor has not been given much study.
Sensitivity	Extent to which test reflects effects of conditions of study.	Alluisi (1967, 1969) Grodsky (1967) Reilly & Cameron (1968) Theologus et al. (1973) Rose (1974) Teichner (1974)	Alluisi (1967, 1969), Grodsky (1967), and Theologus et al. (1973) emphasize sensitivity to effects only to magnitude experienced in operational situation. Reilly & Cameron (1968) define sensitivity as extent to which conditions are likely to influence performance. Teichner (1974), however, discusses it in terms of quickness of detecting dysfunctions. Sensitivity $S = (M_1 - M_2)/\sqrt{SD_1^2 + SD_2^2}$, where M_1 and SD_1^2 are the mean and within cell variance under condition "i" appears a more useful metric for purposes such as Teichner's (1974).
Trainability	Asymptotic levels of performance should be attainable with the selected tasks after a minimum of training except where tasks are selected to measure changes in this function per se.	Alluisi (1967, 1969) Kennedy (1971) Theologus et al. (1973) Rose (1974)	To date studies to insure "asymptotic levels of performance" have not been accomplished for non-learning tasks. Development of tasks to measure different types of learning per se is very lacking though, as Teichner (1974) point out, the most sensitive test will have "minimal practice" as a characteristic. Learning tasks appear to have high potential for future PTBs and should be more fully studied.
Efficiency	Importance of test's contribution with respect to cost, time and effort of implementation.	Reilly & Cameron (1968) Theologus et al. (1973) Teichner (1974)	Contributions of tasks in terms of cost, time and effort appear to be accomplishable by appropriate analysis. The reliability of a task for one minute of study (r_{11}), for example, can be estimated by, $r_{11} = r_{tt}/(t + (1 - t)\, r_{tt})$, where r_{tt} is the observed reliability for t minutes of observation.

REFERENCES

Adams, J. A. Human tracking behavior. Psychological Bulletin, 1961, 58, 1, 55 - 79.

Alluisi, E. A. Optimum uses of psychobiological sensorimotor and performance measurement strategies. Human Factors, 1975, 17 (4), 309 - 320.

Alluisi, E. A. Methodology in the use of synthetic tasks to assess complex performance. Human Factors, 1976, 9 (4), 375 - 384.

Bachrach, A. J. Psychological research: An introduction. (2nd ed.) New York: Random House. 1965.

Bachrach, A. J. & Kennedy, R. S. Psychological performance testing under water and pressure: Problems and prospects. Bethesda, Md.: U.S. Naval Medical Research Institute, 1977. (In press).

Bradley, J. V. Studies in research methodology. III. The persistence of sequential effects despite extended practice. Wright-Patterson Air Force Base, Ohio, MRL Technical Document 62-60, June 1962.

Broadbent, D. Noise, paced performance, and vigilance tasks. British Journal of Psychology, 1953, 44, 295 - 303.

Cross, K. A. & Bittner, A. C., Jr. Accuracy of altitude, roll angle, and pitch angle judgments as a formation of size of vertical contact analog display. Point Mugu, CA: Naval Missile Center, January 1969 (PM-69-2).

Fleishman, E. A., & Hemple, W. E., Jr. The relation between abilities and improvement with practice in a visual discriminatory task. Journal of Experimental Psychology, 1955, 49, 301-312.

Fregly, A. R., & Kennedy, R. S. Comparative effects of prolonged rotation at 10 rpm on postural equilibrium in vestibular normal and vestibular defective human subject. Aerospace Medicine, 36, 12, 1965.

Guedry, F. E., Jr., Kennedy, R. S., Harris, C. S., & Graybiel, A. Human performance during two weeks in a room rotating at three rpm. Aerospace Medicine, 35, 11, 1964.

Graham, C. H. Visual perception. In S. S. Stevens (Ed.) Handbook of experimental psychology. New York: Wiley & Sons, 1951.

Graybiel, A., Kennedy, R. S., Knoblock, E. C., Guedry, F. E., Jr., Mertz, W., McLeod, M. E., Colehour, J. K., Miller, E. F., II, & Fregly, A. R. Effects of exposure to a rotating environment (10 rpm) on four aviators for a period of twelve days. Aerospace Medicine, 36, 8, 1965.

Grodsky, M. A. The use of full scale mission simulation for the assessment of complex operator performance. Human Factors, 1967, 9 (4), 341 - 348.

Kennedy, R. S. Individual differences in auditory vigilance performance on the band-pass ability (B-PA) test: Some theoretical considerations. Presented at Human Factors Society annual meeting, San Francisco, CA, October 1970.

Kennedy, R. S. A sixty-minute task with 100 scoreable responses. Naval Aerospace Medical Center, Pensacola, Florida, NAMI-1045, 1968.

Kennedy, R. S. A performance assessment in exotic environments: A flexible, economical, and standardized vigilance test. Paper presented at the Fifteenth Annual Human Factors Society Meetings, New York, October 1971.

Kennedy, R. S., & Bruns, R. A. Consideration for the utilization of a flexible economical, vigilance test to assess performance in exotic environments. Presented at the October 1975 Aerospace Medical Panel Specialists' Meeting, Ankara, Turkey.

Kennedy, R. S., & Hutchins, C. W. Relationships between physical fitness, endurance, and success in flight training. Naval Aerospace Medical Center, Pensacola, Florida, NAMI-1088, in press, 1971.

Kennedy, R. S., Moroney, W. F., Bale, R. M., Gregoire, H. G., & Smith, D. C. Motion sickness symptomatology and performance decrements occasioned by hurricane penetrations in C-121, C-130, and P-3 Navy aircraft. Aerospace Medicine, 43, 1235 - 1239, 1972.

Kennedy, R. S., Tolhurst, G. C., & Graybiel, A. The effects of visual deprivation on adaptation to a rotating environment. NSAM-918. NASA Order No. R-93. Pensacola, FL: Naval School of Aviation Medicine, 1965.

Kennedy, R. S. PETER for Mentation Mensuration. A point paper, Naval Aerospace Medical Research Laboratory Detachment, New Orleans, LA, December 1977. (In press).

Parker, J. F., Jr. & Fleishman, E. A. Ability factors and component performance measures as predictors of complex tracking behavior. Psychological Monographs, 1960, 74 (16, Whole No. 503).

Parker. J. F., Jr. & Fleishman, E. A. Use of analytical information concerning tasks requirements to increase effectiveness of skilled training. Journal of Applied Psychology, 1961, 45, 295-303.

Parker, J. F., Jr. Use of an engineering analogy in the development of tests to predict tracking performance. The Matrix Corporation. (Office of Naval Research Contract No. ONR-3065(00)). February 1964.

Reilly, R. E. & Cameron, B. J. An integrated measurement system for the study of human performance in the underwater environment. ONR Contract N0014-67-C-0410, December 1968.

Rose, A. M. Human information processing: An assessment and research battery. University of Michigan, Technical Report No. 46.

Rose, A. M. Human information processing: An Assessment and research battery. Ann Arbor, MI., University of Michigan, Doctoral dissertation, 1974, also published as AFOSR-PR-74-1372 (AD-785-411).

Rosenthal, R. Experimental outcome--orientation and the results of the psychological experimentation. Psychology Bulletin, 1963, 61, (6), 405-442.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. Decision processes in perception. Psychological Review, 1961, 68, 301 - 340.

Teichner, W. H. Quantitative models for predicting human visual/perceptual/motor performance. New Mexico State University/Office of Naval Research - Technical Report 74-3, Las Cruces, New Mexico, October 1974.

Theologus, G. C., Wheaton, G. R., Mirabella, A., Brakler, R. E., & Fleischman, E. A. Development of a standardized battery of performance tests for the assessment of noise stress effects, NASA CR 2149, Washington, D. C., 1973.

Thorndike, R. L. Personnel selection: Tests and measurements techniques. New York: Reilly, 1949

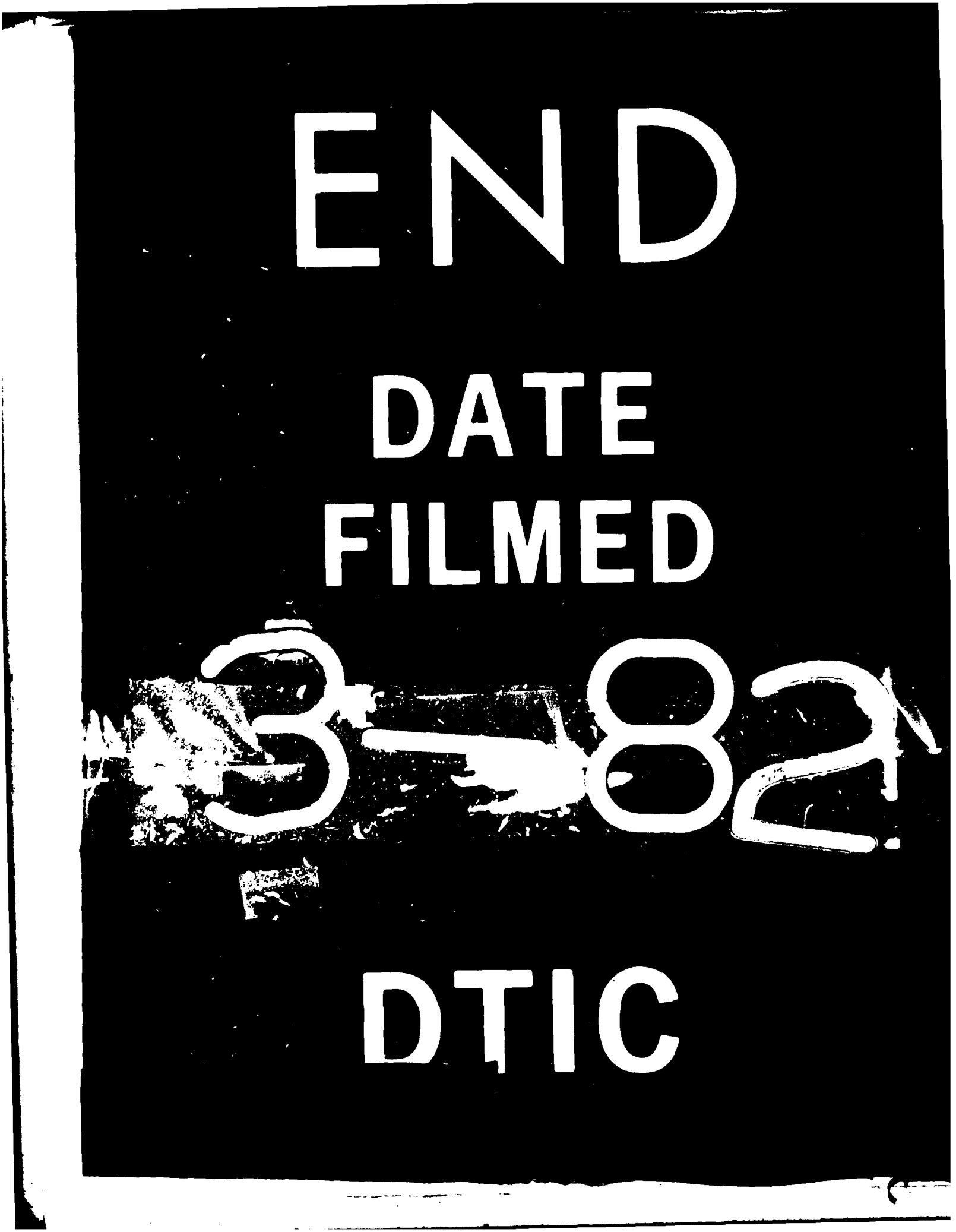
Woodrow, H. The effect of practice on groups of different initial ability. Journal of Educational Psychology, 1938, 29, 268-278(b).

Woodrow, H. The relation between abilities and improvement with practice. Journal of Educational Psychology, 1938, 29, 215-230(a).

Woodworth, R. S. & Schlosberg, H. Experimental psychology. New York: Henry Holt & Co., 1954.

ABOUT THE AUTHOR

Robert S. Kennedy has been an aerospace experimental psychologist since he entered the Navy in 1959. He received an MA in experimental psychology from Fordham University in 1959 and a Ph.D. from the University of Rochester in 1972. His previous military experience includes two tours in the Aerospace Psychology Division at the Naval Aerospace Medical Institute, Pensacola, Florida, where he conducted research on vestibular function, motion sickness, vigilance, and habituation in exotic environments; one tour at the Behavioral Sciences Department at the Naval Medical Research Institute, Bethesda, Maryland, working on the Man-in-the-Sea program; one tour at the Pacific Missile Test Center, Point Mugu, California; and one tour at the Air Development Center, Warminister, Pennsylvania, where he worked mainly on the development, test, and evaluation of airborne weapons systems from the standpoint of human factors engineering. Presently, he is the Officer-in-Charge of the Naval Aerospace Medical Research Laboratory Detachment working on human performance mensuration in unusual environments, specifically ship motion.


END
DATE
FILMED
3-82
DTIC